SOME ASPECTS OF STATISTICAL ECOLOGY

A DISSERTATION SUBMITTED FOR

Master of Philosophy

IN

STATISTICS

BY

NASEEM AHMAD

Under the Supervision of

Prof. Mohammad Zubair Khan

DEPARTMENT OF STATISTICS & OPERATIONS RESEARCH
ALIGARH MUSLIM UNIVERSITY
ALIGARH (INDIA)

1994

DEDICATED
TO
MY PARENTS

SOME ASPECTS OF STATISTICAL ECOLOGY

PREFACE

This dissertation entitled "**Some aspects of Statistical Ecology**" is being submitted to the **ALIGARH MUSLIM UNIVERSITY** for the requirement of the degree of **Master of Philosophy** in Statistics.

To many the title may seem surprising and even strange. To some, ecology is alright but what is this Statistical Ecology. They may not have heard and may not have even fancied any thing like statistical ecology.

The last two decades has seen a sharp rise of interest in ecological problems all over the world. This factor, plus all - prevailing talk of environment - a clean environment necessary for human survival - has been the motivation for us to work in the field of statistical ecology.

"Statistical Ecology encompasses numereous methodologies that deal with the exploration of patterns in biotic Communities". Ecology is a science synthetic in principle, uses a vide variety of methods, particularly the most powerful methods of modern natural sciences namely the mathematical and statistical methods, and thus a new science in the name of statistical ecology has emerged and grown rapidly. " Statistical Ecology" falls within the broader

arena of what is popularly known as mathematical or quantitative ecology, which encompasses both population dynamics and community patterns.

Our dissertation surveys, the various articles in problems of ecology - not encompassing all the aspects but only some of it, because of certain limitations.

This dissertation is organized into four chapters. The first chapter deals with basic concepts and lays the ground work for the subsequent material. Chapter second deals with association and segregation between the species in a discrete and continuous case and also with the effect of Quadrat size. In chapter third, we describe how the abundances of the different species distributed in a community. Some diversity indicies are also discussed. In the fourth chapter, we deal with the probability of discovering a new species and estimate them by the help of parametric and nonparametric estimation.

I like to express my indebtedness and sincere gratitude to my supervisor **Prof. MOHAMMAD ZUBAIR KHAN**, for his guidance and valuable suggestions in writing this dissertation. His great invovlement and sympathetic attitude enabled me to complete this work in due time.

I am extremely grateful to Prof. S.U. KHAN, Chairman, Department of Statistics of Operations Research, for

providing me the necessary facilities. I am also indebted to Prof. A.H. KHAN for his help and co-operation all along.

I express my thanks to Dr. ASAD. R. REHMANI and MRS. NASREEN HUSSAIN, the Centre of Wildlife and Ornithology, for their constant help and encouragement.

I take the opportunity to thanks all the research scholars of the Department, particularly I would like to mention Mr. TARIQ RASHID, Mr. YOUSUF WANI and Mr. ZAHEER KHAN for their constant encouragement that enabled me to pursue studies and write this dissertation.

Finally my thanks go to Mr. MUNIR-UDDIN KHAN in Typing and retyping the material with great sincerity.

Place : **Aligarh**

Dated : **7th May, 1994**

(**NASEEM AHMAD**)

CONTENTS

CHAPTER I

PREIMINARIES AND BASIC RESULTS

CHAPTER I

PRELIMINARIES AND BASIC RESULTS

1.1. INTRODUCTION:

The ecology means "the study of living things in their surroundings. The study of animals and plants in relation to their habit and habitats. By ecology we mean the body of knowledge concerning the economy of nature - the investigation of the total relations of the animal both to its inorganic and to its organic environment.

Ecology concerns itself with the interrelationships of living organisms, plant or animal and their environments; these are studied with a view to discovering the principles which govern the stability of biological communities. That principles exist is a basic assumption - and an act of faith - of the ecologist. His field of inquiry is the totality of the living conditions of the plants and animals under observation, their systematic position,

their reactions to the environment and to each other, and the physical and chemical nature of their surroundings. Ecologists take help from many disciplines.

Statistical ecology encompasses numerous quantitative methodologies that deal with the exploration of patterns in biotic communities. These patterns are of many different types, including the spatial dispersion of a species "within" a community, relationship between many species "within" a community, and relationships among many species "between" communities. Hence, our definition of Statistical ecology falls within the broader arena of what is popularly known as Mathematical or quantitative ecology which encoppasses both population dynamics and community patterns.

Ecology data in community ecology may be viewed as a product of either an experimental or observational approach (Goodall 1970).

An experimental approach presupposes that the community is subject to experimental manipulation. That is, we can divide the community into replicate portions on which various treatments and controls can be imposed. Therefore, any differences detected in measured responses can be attributed to the experimental treatments. On the other hand, using an observational approach, we make measurements on the community over a range of conditions imposed by nature rather than by the researcher. This leaves us with two alternatives: (1) to study different samples obtained at the same time but under different conditions (e.g., phytoplankton sampling of inshore and off - shore waters of a lake) or (2) to study samples at the same place but at different times (e.g. of offshore phytoplankton taken during the summer and winter).

1.2. Ecological Sampling:-

The various stages of an observational study are shown in Figure 1.1. The first step involves a clear definition of the aims of the study.

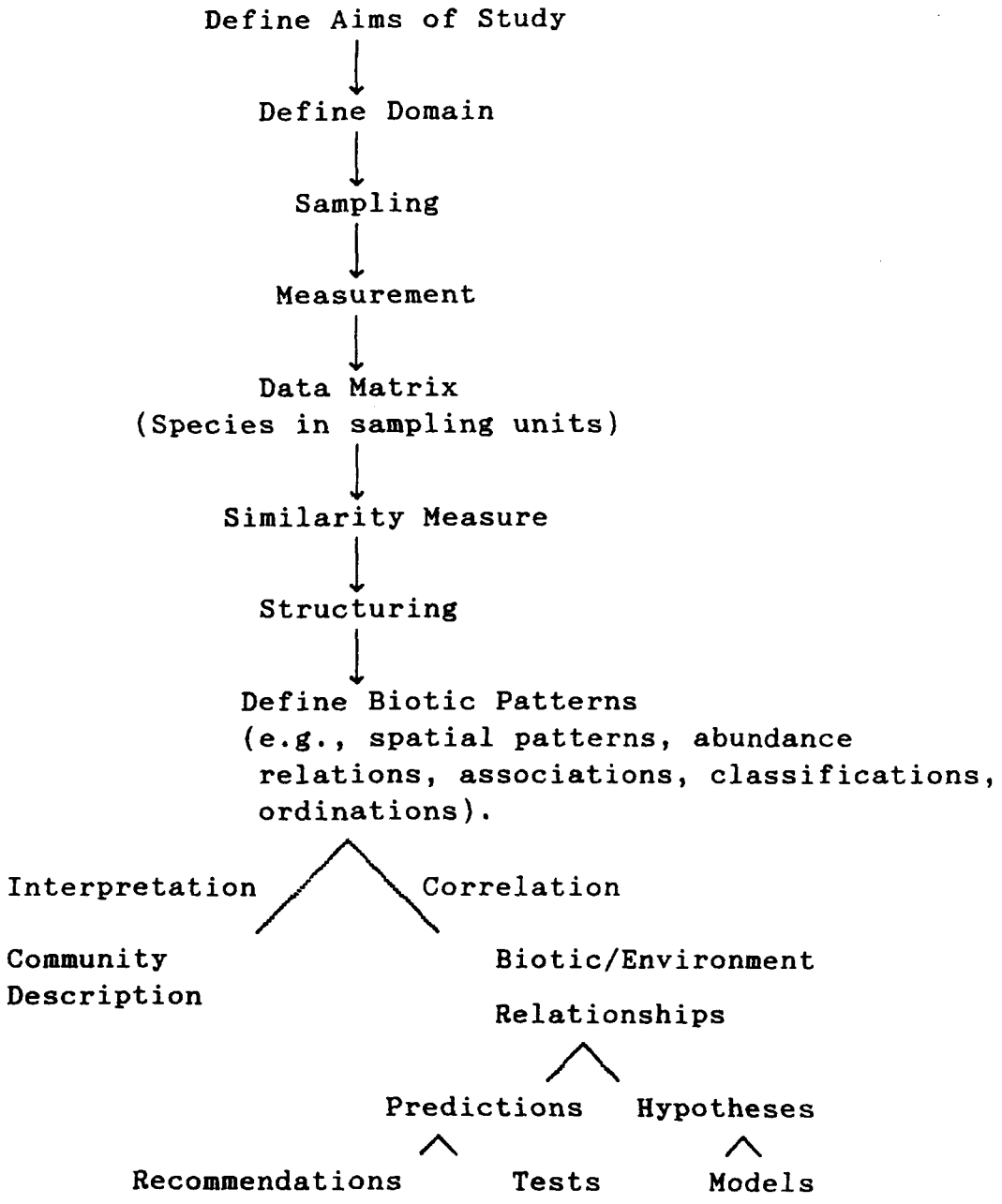


FIGURE 1.1. STAGES OF AN OBSERVATIONAL ECOLOGY APPROACH (AFTER NOY-MEIR 1970).

A successful sampling scheme involves the selection of an appropriate sampling unit (SU). Common sampling units used in ecology include quadrats, leaves of a plant, light traps, soil cores, pit traps individual organisms, and belt transects (Table 1.1) Some SUs occur naturally (e.g. plant leaves), while others are arbitrarily defined (e.g., quadrats). For example using a $1m^2$ SU, a sample might consist of 20 such SUs randomly located throughout the study area.

Once the sampling procedure and the choice of SU have been made, specific measurements (e.g., presence-absence, biomass, density) are taken on the species of interest in the community. These data are then tabulated into an ecological data matrix, which is a convenient method of summarizing large data sets and is the basic unit that we subject to analyses. The data matrix is a rectangular display of the measurements taken in each S.U. There are two basic types of data matrices depending on the purpose of study.

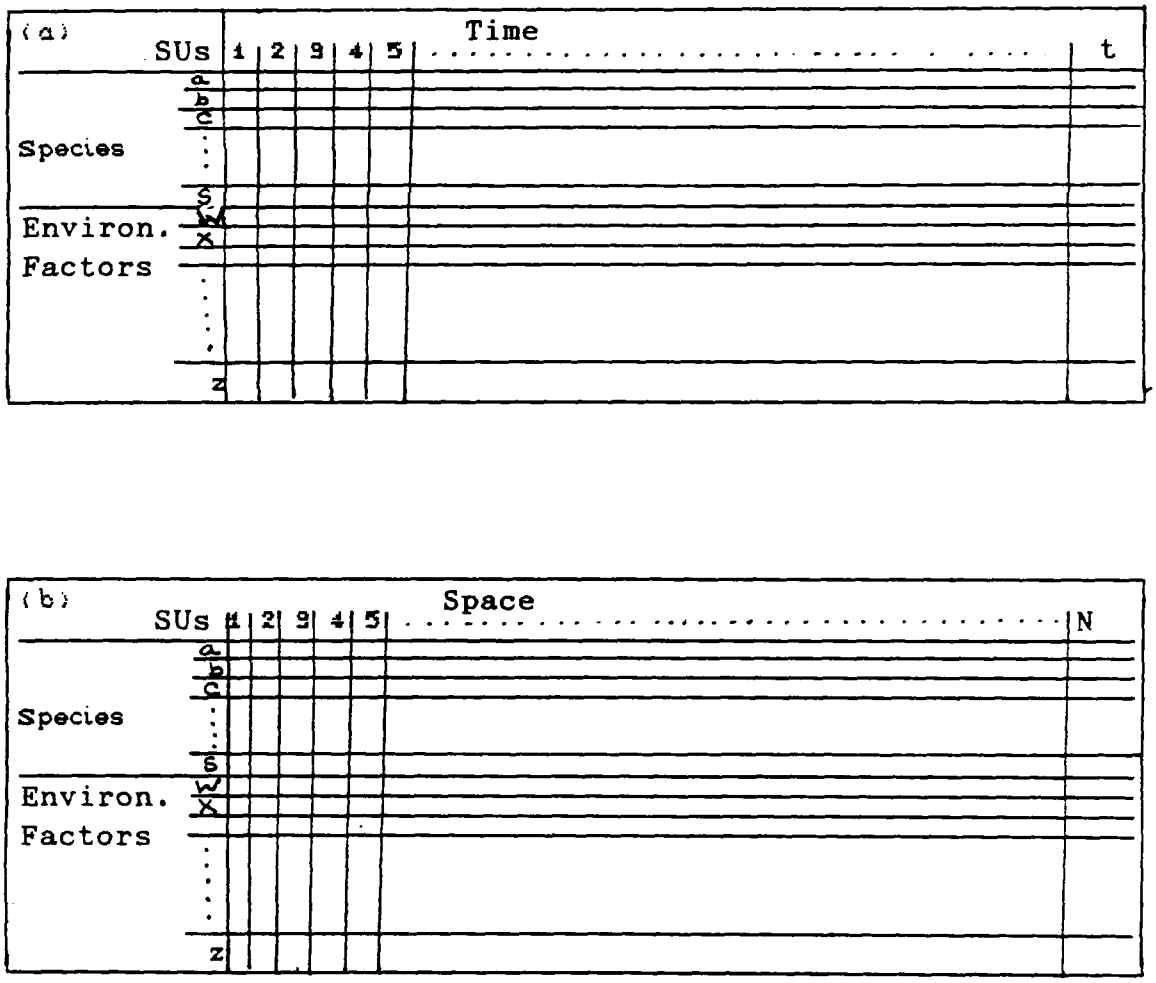


Figure 1.2. Two basic types of ecological data matrices (a) species and environmental factor data measured in one location through time (a total of t observations, the SUs) and (b) species and environmental factor data measured on N SUs over space, that is, at different locations in the landscape.

First, in studies dealing with temporal dynamics (e.g., community succession), the data matrix represents made on species through time (Figure 1.1a). The time interval depends on the specific purpose of the study. Environmental factors (e.g., soil water content, soil pH, soil temperature) are often simultaneously collected in the SUs. The second type of data matrix deals with measurements taken on a number of SUs distributed over space (Figure 1.1b). The actual spatial distribution of the SUs is determined by the experimental design (e.g., random placement of quadrats).

1.3. SPATIAL PATTERNS ANALYSIS:-

The spatial pattern of plants and animals is an important characteristic of ecological communities. This is usually one of the first observations we make in veiwing any community and is one of the most fundamental properties of any group of living organims. (Connell 1963).

Three basic types of patterns are recognized in communities: random, clumped, and uniform (Figure 1.3). Random patterns in a population of organisms imply environmental homogeneity and/or nonselective behavioural patterns. On the other hand, nonrandom patterns.

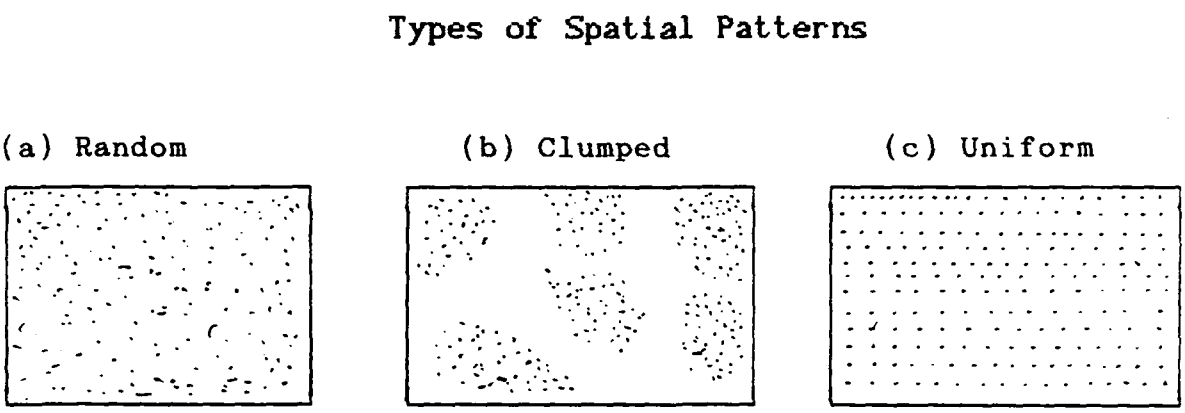


Figure 1.3. These types of spatial pattern: (a) random, where all individuals are located independently of each other; (b) clumped, where individuals tend to be located together in clusters; and (c) uniform where individuals are regularly spaced.

(Clumped and uniform) imply that some constraints on the population exist. Clumping suggests that individuals are aggregated in more favorable parts of the habitat; this may be due to gregarious behavioUr, environmental heterogeneity, reproductive

mode and so on. Uniform dispersions result from negative interactions between individuals, such as competition for food or space. Of Course, detecting a pattern and explaining its possible causes are separate problems.

Hutchinson (1953) was one of the first ecologists to consider the importance of spatial patterns in communities and identified various causal factors that may lead to patterning of organisms (1) vectorial factors resulting from the action of external invironmental forces (e.g., wind, water currents, and light intensity); (2) Reproductive factors attributable to the reproductive mode of the organisms (e.g., cloning and progeny regeneration); (3) social factors due to innate behaviors (e.g., territorial behavior); (4) Coactive factors resulting from intraspecific interaction (e.g., competition); and (5) Stochastic factors resulting from random variation in any of the preceding factors.

The relationships between the mean and variance of the number of individuals per SU is influenced by the underlying pattern of dispersal of the population. We can now define the three basic types of patterns and their variance-to-mean relationships, where σ^2 = variance and μ represents the mean:

1. Random pattern : $\sigma^2 = \mu$
2. Clumped pattern $\sigma^2 > \mu$
3. Uniform pattern : $\sigma^2 < \mu$

There are certain statistical frequency distributions that, because of their variance to mean properties have been used as models of these types of ecological patterns.

(1) The poisson distribution ($\sigma^2 = \mu$) for random patterns

(2) The negative binomial ($\sigma^2 > \mu$) for clumped patterns

(3) The positive binomial ($\sigma^2 < \mu$) for uniform patterns.

While these three statistical models have commonly been used in studies of spatial pattern, it should be recognized that other statistical distributions might be equally appropriate (Pielou) 1977.

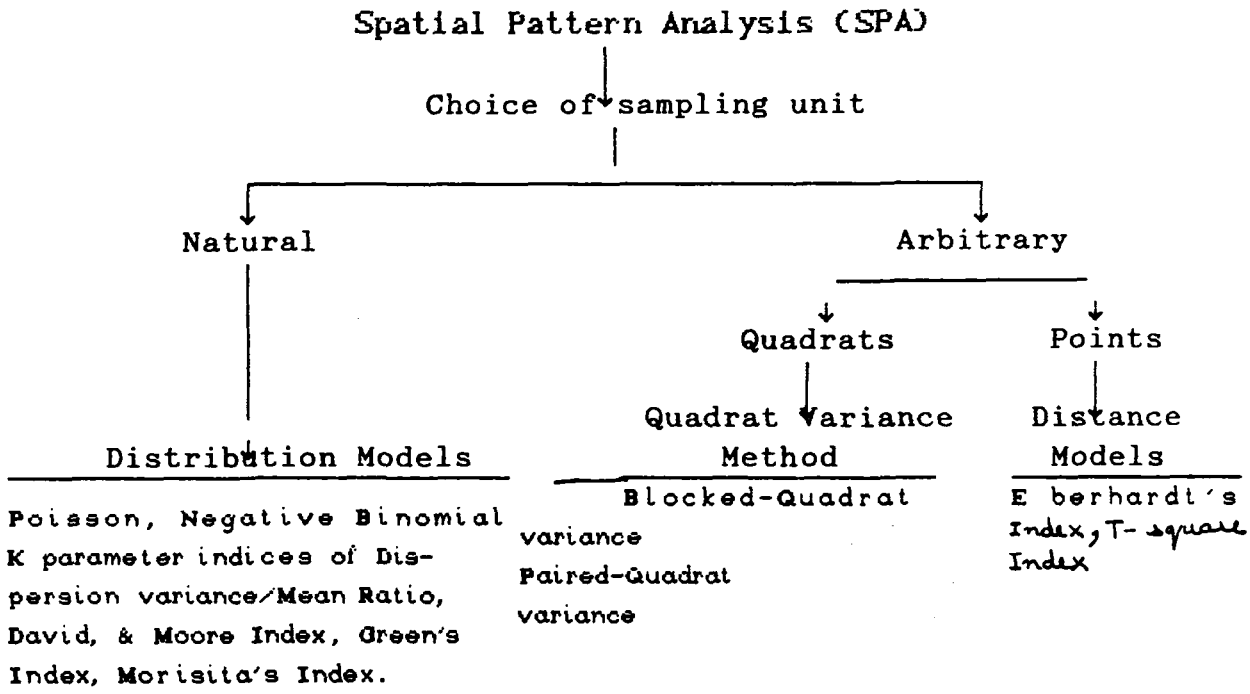


Figure 1.4 Types of SPA models following choice of SU

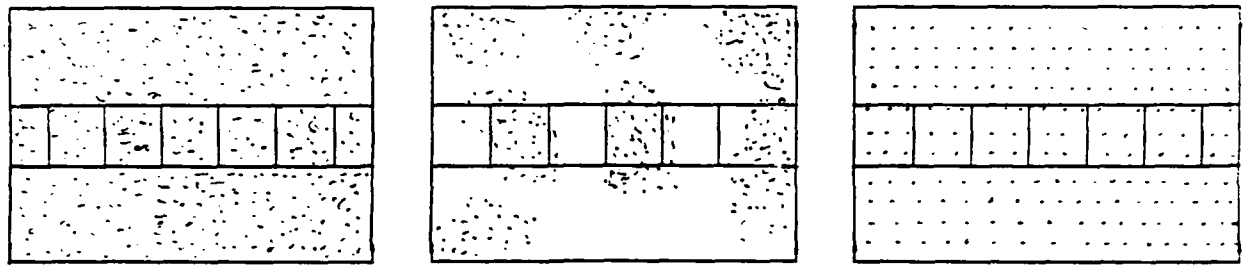
1.4. Quadrat: Refers to an area, usually of vegetation, randomly selected for study. It is normally square in shape (hence the name). The ideal size of a quadrat has been the smallest size that has same number of species as would be contained in a larger one. In a point quadrat sampling has been done at the points of a square grid covering the quadrat.

1.4.1. Quadrat-Variance Methods:

We are concerned with the spatial pattern of the individuals of species that are found continuously across a community (e.g., trees in a forest). With continuous or nondiscrete habitats, some type of arbitrary SU must be chosen to obtain a sample and consequently results may be influenced by the size and shape of the SU chosen. To address this problem, methodologies have been developed that allow us to examine the effect that varying the size of the SU has on the detection of some underlying spatial pattern. Collectively, those are called quadrat-variance methods.

When the dispersion of individuals of a species is continuous over a study area (e.g., a grass species across a grassland community), arbitrary SUs must be used to obtain a sample. As an example, consider the use of a belt or grid of contiguous (abutting) quadrats positioned and observed within such a community as depicted in Figure 1.5. Note that if the spatial pattern of the individuals is random, observations of the number

of individuals per quadrat (and, of course, the mean and variance) will be quite different from those obtained from either the clumped or uniform spatial patterns. The methods are based on data obtained from such sample of contiguous quadrats and may be used to address hypothesis concerning the spatial patterning of inidividuals in a community.



(a) Random (b) Clumped (c) Uniform

Figure 1.5: Positioning of contiguous quadrats (a belt transect) in a community where individuals ar (a) randomly dispersed (b) clumped, or (c) uniformly dispersed over the sample area.

Whenever arbitrary SUs (quadrats) are used in sampling, it is important to be cognizant of the influence that the size and shape of the quadrat might have on the results. For the clumped distribution shown in figure (1.5b) it is obvious that the number

of individuals per quadrat will be greatly influenced by a doubling of the quadrat size. On the other hand, doubling the quadrat size is not a problem when the distribution is random (e.g. Figure 1.5a) the expected number of individuals per quadrat (regardless of size, as long as all are of the same size) is the same thoughout a random population, and the frequency distribution of the number of individuals per quadrat will always follow a Poisson series (Pielou 1977). Quadrat-variance methods are based on examining the changes in the mean and variance of the number of individuals per SU over a range of different SU sizes.

1.5. Species Affinity:

Ecological communities are composed of a number of coexisting species. Some Communities may have a large number of species (e.g., a tropical forest); others may have just a few (e.g., a polluted river). We know that some empirical models for quantifying the relationships between the total number of species in a community and some measure of their abundances (e.g. total numbers). Here we are interested in examining the affinities of

coexisting species. How do coexisting species utilize common resources ?

Consider, for example, a species-rich lake that has four dominant fish, all about the same size. Are they in direct competition for food and space ? Do some species feed exclusively in the surface waters, while others feed on the lake bottom ? When we spatially locate species A, are we likely more often than not to find species B there as well ? In a broad sense, we can define such interspecific interactions as the degree of affinity between species.

One measure of affinity is the degree to which species overlap in their utilization of common resources. This overlap is defined in terms of various portions of the species niche that is should by other species. Niche studies are based on such species attributes as diet, microhabitat preference, and timing of activities (e.g. foraging).

In the interspecfic association we are concerning only with measuring how often two species are found together by examining if the occurence of the species [in a series of sample units (SUs)] is greater than or less than what would be expected if they were independent. If either positive or negative association is detected, we can measure the strength of this association with indices.

Association is based solely on presence/absence data. If a sample contains quantitative measures of species abundance, we can determine the covariation in abundances between species. This may lead to questions concerning species affinities. For example, if the abundance of one species always decreases when the other species increases, is there some type of causal negative interaction ?

1.6.SOME STATISTICAL DISTRIBUTIONS

1.6.1. Logarithmic Series Distribution:

The random variable x has a logarithmic series distribution if

$$(1)\quad P_r(X=K) = \alpha\theta^k/K \qquad (k=1,2,\ldots,..0<\theta<1)$$

Where $\alpha = -[\log(1-\theta)]^{-1}$ The probabilities are the terms in the series expansion of $-\alpha\log(1-\theta)$.

The moment generating function of x is

$$(2)\quad E(e^{tx}) = \log(1-\theta e^t)]/[\log(1-\theta)].$$

The probability generating function is

$$(3)\quad \alpha\sum_{j=i}^{\alpha}\left[\theta^j/j\right]t^j = [\log(1-\theta t)]/[\log(1-\theta)].$$

The rth factorial moment is

$$(4)\quad \mu_{(r)} = E\{X^{(r)}\} = \alpha\theta^r\sum_{k=r}^{\infty}(k-1)(k-2)\ldots.(k-r+1)\theta^{k-r}$$

$$\mu_{(r)} = \alpha\theta^{r} \frac{d}{d\theta^{r-1}} \left[\sum_{k=1}^{\infty} \theta^{k-1} \right]^{r-1}$$

$$\mu_{(r)} = \alpha\theta^{r}(r-1)!(1-\theta).^{-r}$$

The moment ratios $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \frac{\mu_4}{\mu_2^2}$ both tend to ∞ as θ tends to 0 or as θ tends to 1, with

$$\lim_{\theta \to 0} (\beta_2/\beta_1) = 1; \quad \lim_{\theta \to 1} (\beta_2/\beta_1) = \frac{3}{2}.$$

1.6.2. Poisson Distributions:

A random variable X is said to have a Poisson distribution with parameter θ if

(1) $P_r [X=K] = e^{-\theta}\theta^k/k!$ $(k=0,1,2\ldots;\theta>0)$

This distribution is the limit of a sequence of binomial distribution with

$$P_{k,N} = Pr[X=K] = \begin{cases} \binom{N}{K} p^{k}(1-P)^{N-K} & (\text{for } k=0,1..N) \\ = 0 & (\text{for } k>N) \end{cases}$$

in which N tends to infinity, and p tends to zero but Np remains equal to θ. It can be established by direct analysis that

$$(2)\ \lim_{\substack{N\to\infty \\ N_p\to\theta}} \sum_{w} P_{K,N} = \sum_{w} e^{-\theta}\theta^{k}/K!$$

Where $\sum_{w}$ denotes summation over any (finite or infinite) subset w of the non-negative integers 0,1,2 ....

1.6.3. Negative Binomial Distribution:

The negative binomial distribution with parameters N,P is defined as the distribution of a random variable, x, for which

$$(1)\ P_{r}[x=k] = \binom{N+K-1}{N-1} (P/Q)^{K}(1-P/Q)^{N} \quad (K=0,1,2,....)$$

The parameter M=NP (the expected value) is often used instead of P, giving the form

$$(2)\quad P_r\ [X=K] = \binom{N+K-1}{N-1}\left(M/N\right)^k\left(1+M/N\right)^{-N+K} \quad (k=0,1,2\ldots).$$

Note that there is a non-zero probability for x taking any specified non-negative integer value, as in poisson distribution, but unlike the binomial distribution. N need not be an integer. When N is an integer, the distribution is sometimes called the Pascal distribution.

1.6.4. Log Normal Distribution

If there is a number θ, such that Z = log(X-θ) is normally distributed, the distribution of x is said to be lognormal. For this to be the case it is clearly necessary that x can take any value exceeding θ, but has zero probability of taking any value less than θ. The name 'lognormal' can also be applied to the distribution of x if log(θ-X) is normally distributed, X having zero probability of exceeding θ. However, since replacement of x by -x, (and θ by -θ) reduces this situation to the first, we will

consider only the first case.

The distribution of x can be defined by the equation.

(1) $U = \gamma + \delta \log (x-\theta)$

Where U is a unit normal variable and γ, δ and θ are parameters. From (1) it follows that the probability density function of X is

(2) $p_x(X) = \delta[(X-\theta)\sqrt{2\pi}]^{-1} \exp\left[-\frac{1}{2}\{\gamma+\delta \log (x-\theta)\}^2\right] .. (X>\theta)$.

1.6.5. Chi-Square Distribution:

The square of a standard normal variate is known as a chi-square variate with 1 d.f.

Thus if $X \sim N(\mu,\sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$ is N (0,1)

and $Z^2 = (\frac{X-\mu}{\sigma})^2$ is a chi-square variate with 1.d.f.

In general If X_i (i=1,2,.. ..n) are n independent normal variates with mean μ_i and variance σ_i^2 (i=1,..n)

Then

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{X_i - \mu_i}{\sigma_i}\right]^2$$ is a chi-square variate with n.d.f.

The probability density function of χ^2 with n degree of freedom is

$$dp(\chi^2) = \frac{1}{2^{\frac{n}{2}}\left[\frac{n}{2}\right.} e^{-\chi^2/2} (\chi^2)^{\frac{n}{2}-1} d\chi^2, \qquad 0<\chi^2<\infty$$

Remarks:-

(1) If χ^2 is a chi-square variate with n.d.f., then $\chi^2/2$ is a Gamma variate with parameter $\frac{n}{2}$. Symbolically, if $\chi^2 \sim \chi^2_n$ then $\chi^2/2 \sim \gamma\ (\frac{n}{2})$ variate.

(2) Since the probability function $f(\chi^2)$ does not involve any population parameter, χ^2-test is sometimes considered to be a non-parametric test.

1.7. Criteria of Estimation

1.7.1. Unbiasedness:

Unbiasedness is a property associated with finite n. A statistics $\hat{\theta}_n = \hat{\theta}_n (x_1, x_2, \ldots . x_n)$ is said to be an unbiased estimate of parameter θ if

$$E(\hat{\theta}_n) = \theta.$$

1.7.2. Consistency:

Let $\hat{\theta}_n$ be an estimator of θ based on a sample of size n. Then $\hat{\theta}_n$ is a consistent sequence of estimator of θ (or $\hat{\theta}_n$ is consistent for θ, briefly) if

$$\hat{\theta}_n \xrightarrow{P} \theta \text{ as } n \longrightarrow \infty$$

i.e. for every $\in > 0$,

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \in) = 0.$$

or equivalently, for every $\in > 0$, $\eta > 0$.

$$P(|\hat{\theta}_n - \theta| < \in) > 1-\eta, \quad n \geq N$$

Where N is some very large value of n.

1.7.3. Efficiency:

If $\hat{\theta}_1$ is the most efficient estimator with variance, V_1 and $\hat{\theta}_2$ is any other estimator with variance V_2 then the efficiency E of $\hat{\theta}_2$ is defined as $E = \frac{V_1}{V_2}$

obviously, E can not exceed unity.

1.7.4. Sufficient Estimator:-

An estimator t_n is said to be sufficient for estimating a population parameter θ, if it contains all the informations in the samples about the parameter.

1.7.5. Minimum Variance Unbiased Estimator (MVUE)

If a statistic $t = t(x_1 x_2, \ldots . x_n)$ based on a sample of size n is such that

(i) t is unbiased and

(ii) It has smallest variance among the class of all unbiased estimator of θ, then t is called minimum variance unbiased estimator (M.V.U.E.) of θ. More precisely t is a M.V.U.E. of θ if $E(t) = \theta$

and $V(t) \leq Var(t')$

Where t' is any other unbiased estimator of θ

That is

$$E(t') = \theta$$

1.7.6. Non parametric Methods:

Most of the test require specific assumption about the population or the populations sampled. In most cases we assume that the populations sampled are normal, sometimes we assumed that their standard deviations are known or are known to be equal. These type of tests are known as parametric tests. There are many situations in which the required assumptions can not be met, alternative techniques, which have become known as nonparametric methods, have been developed. This term is used, some what loosely to include distribution free methods where we make no assumptions about the population, except that they are continuous. The non-parametric methods can be used under more general conditions than the standard techniques. In addition, they are easy to explain and easy to understand, But the drawback is they are less efficient.

CHAPTER II

SPATIAL RELATIONS OF TWO OR MORE SPECIES

CHAPTER II

SPATIAL RELATIONS OF TWO OR MORE SPECIES

2.0. INTRODUCTION:

The spatial pattern exhibited by a single species within a limited area is often worth examining for its own sake. The factors controlling and determining pattern, however are likely to affect many species rather than just one, and much may be learned by investigating the way in which species are associated with one another. If two co-occurring species are affected by the same environmental factors, or if they have some effect, either favorable or unfavorable. On each other, their patterns will not be independent; the species will be associated either positively or negatively. Association or the lack of it among pairs and groups of species is therefore of obvious ecological interest. As in the study of pattern in one-species population it is desirable to consider separately those species that occupy discrete habitable units (e:g, pest insects in fruits), and those that may occur anywhere throughout an extended space or continuous (e.g. plankton organisms in a volume of water, plants in a meadow).

We confine attention to organisms in discrete units and begin by discussing the association of a single pair of species. The far more difficult problem of investigating association in a group of more than two species is mentioned briefly in next section.

2.1. ASSOCIATION BETWEEN TWO SPECIES AND CONSIDERATION OF QUADRAT SIZE.

We are concerned here with measuring how often two species are found in the same location. This affinity (or lack of it) for coexistence of two species is referred to as interspecific association. In general, an association between two species exists because: (1) both species select or avoid the same habit-at or habitat factors; (2) They have the same general abiotic and biotic environmental requirements; or (3) on or both of the species has an affinity for the other, either attraction or repulsion.

The study of species association involves two distinct components. The first is a statistical test of hypothesis that two species are associated or not at some predetermined probability level. The second is a measure of the degree of strength of the association.

We are usually interested in testing for association between two species. Assume that we are examining a sample of N discrete units collected at random from a large population of possible units. Let the two species bring studied be labeled species A and species B; for each unit note whether it contains species A alone, species B alone, both species, or neither species. The quantity of each species in each unit is disregarded; we record only presences and absences. The observed frequencies can then be set out in the form of a 2 x 2 table:

		Species B		
		Present	Absent	
Species A	Present	a	b	m = a + b
	Absent	c	d	n = c + d
		r = a + c	S = b+d	N= m+n = r+s

The usual approach is to carry out a χ^2- test and let it go at that. But this is a slovenly approach, and it is important to realize that there are two entirely different questions we could ask on being confronted with a table such as this (see Peason, 1947).

Question 1. Among the N units examined do species A and species B occur independently of each other ?

Question 2. In the population as a whole are the two species independent of each other ?

Since the table's marginal total are fixed. The question 1 becomes: for the given marginal totals what are the probabilities of the various possible sets of cell frequencies (or partitions of N) ? Is the particular set of cell frequencies we have observed consistent with the hypothesis of independence ? For given N, M and r, the conditional probability that 'a' of the units will contains both species is

$$\Pr(a/N,m,r) = \frac{m!\ n!\ r!\ s!}{a!\ b!\ c!\ d!\ N!};$$

that is, a has a hypergeometric distribution. To see this note that the number of ways of choosing m units out of N to contain A is $\binom{N}{m}$; similarly, the number of ways of choosing r units to contain species B is $\binom{N}{r}$. Thus the number of arrangements that would give rise to the observed marginal totals is $\binom{N}{m}\binom{N}{r}$.

The number of different ways of partitioning N to produce the observed cell frequencies a,b,c, and d is N!/(a!b!c!d!). Therefore

$$\Pr(a|N,m,r) = \frac{N!/(a!b!c!d!)}{\binom{N}{m}\binom{N}{r}}$$

$$= \frac{m!\ n!\ r!\ s!}{a!\ b!\ c!\ d!\ N!}$$

In this way use may calculate the probabilities for all the different sets of cell frequencies that give rise to the observed marginal totals.

We resume discussion of this case after considering how the second case (Question 2) differs from it. In asking Question 2, it is no longer assumed that the marginal totals are fixed. When a sample of N units is taken at random from a large population of units, not only are the cell frequencies free to vary but so also are their pair-wise sums, the marginal totals. To determine the probability of obtaining any particular 2 x 2 table we must argue as follows. Let P(A) denote the probability that a unit will contain species A and $P(\bar{A}) = 1-P(A)$, the probability that unit will lack it. The probabilities p(B) and $p(\bar{B}) = 1-p(B)$ are defined likewise for species B. Any unit must belong to one of four classes, AB, $A\bar{B}$, $\bar{A}B$ or $\bar{A}\bar{B}$, and on the null hypothesis of independence of the species. We must have

$$p(AB) = p(A)\,p(B), \quad p(A\bar{B}) = p(A)\,p(\bar{B})$$

$$p(\bar{A}B) = p(\bar{A})\,p(B), \quad p(\bar{A}\bar{B}) = p(\bar{A})\,p(\bar{B})$$

The probability Pr (a,b,c,d) of obtaining the observed cell frequencies a,b,c and d in a sample of N units is then a term

from a multinomial distribution; it is given by the coefficient of $Z_1^a Z_2^b Z_3^c Z_4^d$ in the expansion of the probability generating function

$$[p(AB)Z_1 + p(A\bar{B})Z_2 + p(\bar{A}B)Z_3 + p(\bar{A}\bar{B})Z_4]^N;$$

that is

$$\Pr(a,b,c,d) = \frac{N!}{a!\,b!\,c!\,d!}[p(AB)]^a[p(A\bar{B})]^b[p(\bar{A}B)]^c[p(\bar{A}\bar{B})]^d$$

$$\Pr(a,b,c,d) = \frac{N!}{a!\,b!\,c!\,d!}[p(A)]^{a+b}[p(B)]^{a+c}[p(\bar{A})]^{c+d}[p(\bar{B})]^{b+d}$$

$$= \frac{N!}{m!\,n!}[p(A)]^m[1-p(A)]^n$$

$$x \frac{N!}{r!s!}[p(B)]^r[1-p(B) \frac{m!n!r!s!}{a!b!c!d!N!}$$

$$= b(m|p(A), N) \times b(r|p(B), N) \times \Pr(a|N,m,r)$$

Here the binomial term b (m/p(A)N) denotes the probability

that in N trials an event whose probability is p(A) will occur m times; $b(r|p(B), N)$ denotes like wise and $Pr(a|N, m,r)$ is conditional probability we found in answering Question 1, namely, the probability of observing the partition of N into the parts a,b,c and d given that a + b = m and a + c = r.

The probability of obtaining an observed 2 x 2 table thus depends on whether the table is assumed to have pre-assigned marginal totals, in which case it represents what Bernard (1947) has called a doubly restricted double dichotomy; this is the assumption made when Question 1 is asked. Or whether the marginal totals as well as the cell frequencies are treated as random variates, giving a table that is an unrestricted double dichotomy; this is the assumption made when question 2 is asked.

2.1.1. Testing the Association in a Sample (Question 1)

If the empirical table is treated as a doubly restricted double dichotomy, we may calculate $Pr(a|N,m,r)$ for all the possible values of a that could arise, subject to the restriction that the marginal totals are fixed. Denote the minimum and maximum possible values of a by a (min.) and a (max).

Now suppose that the observed frequency a of the event AB (i.e, the observed number of joint occurrences of species A and B) is greater than its expectation E(a):that is $E(a) < a \leq a\ (max)$. This leads us to believe that there may be significant positive association between the species. Then the probability of observing a deviation from expectation as great as or greater than a- E(a), and in the same direction is

$$P_{upper} = \sum_{i=a}^{a(max)} Pr\ (i \mid N,m,r).$$

Thus P_{upper} is the appropriate probability for a one-tail test for positive association. It is the probability, on the null hypothesis of independence of obtaining evidence for positive association as strong as or than that observed.

Likewise, if $a(min) \leq a < E(a)$, so that the data suggest negative association, the probability required for a one-tail test is

$$P_{lower} \sum_{i=a\ (min)}^{a} Pr\ (i \mid N,m,r).$$

To do a two-tail test we must sum the probabilities of obtaining a deviation as great as or greater than that observed in either direction. Thus if the deviation of the observed a from expectation, $| E(a) - a |$, is x, the probability for the two-tail test is

$$\left\{ \sum_{i=a(min)}^{E(a)-x} + \sum_{i=E(a)+x}^{a(max)} \right\} Pr(i | N,m,r)$$

This exact test is easily done with the help of tables such as those of Finney et al (1963) and Bennett and Horst (1966), provided both column totals (or both row totals) are $\leq$ 50. Outside the range of these tables we may make use of the fact that the the distribution of a tends to normality. As already remarked, a is a hypergeometric variate. Its mean and variance are

$$E(a) = \frac{rm}{N} \quad \text{and} \quad Var\ (a) = \frac{mnrs}{N^2(N-1)}$$

So

$$X = \frac{a - E(a)}{\sqrt{Var(a)}}$$

is a standardized normal variate and Normal tables may be used to judge significance. Either a one-tail or a two-tail test may be done.

For large N it is permissible to substitute N for N-1 in the denominator of Var(a). Then X becomes

$$X = \frac{\sqrt{N}\,(ad - bc)}{\sqrt{m\,n\,r\,s}} \quad \text{and} \quad X^2 = \frac{N(ad - bc)^2}{mnrs}$$

We see that X^2, being the square of a standardized normal variate, has the χ^2-distribution with one degree of freedom.

Since the continuous χ^2-distribution is being used to approximate a discrete distribution, it is desirable to make a continuity correction. In calculating X^2 this is done by subtracting $\frac{1}{2}$ from the two observed frequencies that exceed expectation and adding $\frac{1}{2}$ to the two frequencies that fall short

of expectation. Then

$$X^2(\text{Corrected}) = \frac{\left[\,|ad - bc| - N/2\,\right]^2 N}{m\ n\ r\ s}$$

This ensures a closer approximation of the χ^2-integral to the sum of the tail terms of the true, discrete distribution X^2.

It will be seen that the expression here denoted by X^2 is the one often described as χ^2. However, for a function of the observed cell frequencies (in other words, a sample statistic) it is preferable to use the noncommital symbol X^2; the symbol χ^2 should be reserved for the theoretical variate with the χ^2-distribution (see Cochran, 1954).

The foregoing arguments explain why we may use the χ^2-test as an approximation to the exact test appropriate to a doubly restricted double dichotomy. However, ecologists who use the test should never lose sight of the fact that a χ^2-test is automatically two-tailed. so if we sometimes use χ^2-test and at other times (because of low observed frequencies) the exact test, the two-tail form of the exact test should be used. Otherwise

the results are not comparable.

2.1.2 Testing the Association in a population

We come to question 2; that is, we wish to know whether the data yielded by a sample could have come from a population in which the two species are independent. The desired probability is a sum of terms of the form of Pr (a,b,c,d), in which hypergeometric probabilities are wieghted with binomial probabilities. The χ^2-test makes no allowance for the binomial terms and although it is the bet test, the calculated tail probabilities are greater than their true values. This may lead to acceptance of the null hypothesis of independence when it should be rejected -a type II error. At the same time the risk of asserting that there is true association when there is not -a type I error - is reduced.

For sufficiently large samples the error introduced is usually negligible.

2.1.3. Measurements of Association

Besides testing we may wish to measure the strength of the association between two species. Suppose species B occurs in

more of the units than does species A. Then positive association would be as great as possible if A was never found in the absence of B, though there would perforce be some units in which B was found without A. This degree of association can be called complete (see Kendall and Stuart, 1967).However, we might choose to assert that the association was "as great as possible" only when neither species ever occurred without the other. This is called absolute association requires that either b or c (not necessarily both) be zero. For absolute association we must have both b = 0 and c = 0; then m = r = a and n = s = d. Depending on whether we want the coefficient of association to be +1 when the association is complete or absolute, We can use the coefficient Q or V defined as follows (Yule 1912).

$$Q = \frac{ad - bc}{ad + bc}$$

then Q = 1 when either b = 0 or c = 0. or

$$V = \frac{ad - bc}{+(mnrs)^{1/2}}$$

Then $V = \pm 1$ only if $mnrs - (ad - bc)^2 = 0$, but since $m\ n\ r\ s - (ad - bc)^2 = 4abcd + a^2(bc + bd + cd) + b^2(ac + ad + cd) + c^2(ab + ad + bd) + d^2(ab + ac + bc)$

the expression on the right vanishes only if two of the cell frequencies are zero. We can exclude from consideration cases in which the two zeros occur in the same row or the same column, for there would be nothing to test. This leaves either b = 0 and c=0, for which $V = \pm 1$, or a = 0 and d = 0, for which V = -1.

Both coefficients are zero when the association is nil, that is, when observed and expected frequencies are equal, for then

$$a - E(a) = \frac{(ad - bc)}{N} = 0$$

The sampling variance of Q is

$$\text{Var}\ (Q) = \frac{(1 - Q^2)^2}{4} \left\{ \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \right\}$$

The derivation is given in Kendall and Stuart (1967). We shall not consider Q further here. The fact that its use precludes any distinction between complete and absolute association makes it unsuitable as a measure of ecological association. To see this consider the two tables

		Species B		
		+	-	
Species A	+	80	80	160
	-	0	15	15
		80	95	175

		Species B		
		+	-	
Species A	+	80	0	80
	-	0	15	15
		80	15	95

They are strikingly different. In the population tabulated on the left, although all the 80 units containing B also contain A, there are in addition 80 units with only in them. By contrast in the population on the right neither species occurs without the other. For both tables Q = 1, whereas V = 1 only for the Right-hand table; for the left-hand table V = 0.281. Any ecologist would assert that the association shown by the table on the right was by for the greater and V is therefore preferable to Q in ecological contexts.

Two other points to mention about V are the following: (a) $V^2 = \frac{X^2}{N}$, Where X^2 is the test statistic defined before; V^2 is known as the mean- square contingency of the 2 x 2 table. (b) V is a correlation coefficient. Assign to each unit a pair of values (x,y) with

$$x = \begin{cases} 1 \text{ when species A is present} \\ 0 \text{ when species A is absent} \end{cases}$$

$$y = \begin{cases} 1 \text{ when species B is present,} \\ 0 \text{ when species B is absent,} \end{cases}$$

Then $$\text{Cov } (x,y) = \frac{a}{N} - \frac{mr}{N^2} = \frac{ad - bc}{N^2},$$

$$\text{Var}(x) = \frac{mn}{N^2} \text{ and } \text{Var}(y) = \frac{rs}{N^2},$$

and therefore

$$V = \frac{ad - bc}{\sqrt{mnrs}} = \frac{\text{Cov } (x,y)}{[\text{Var}(x)\ \text{Var}(y)]^{1/2}}$$

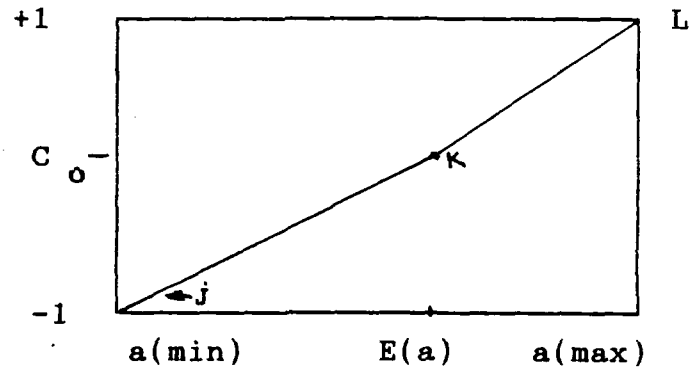
In other words, V is the correlation coefficient between x and y. An estimate of its sampling variance, derived by Yule (1912), is given by

$$\text{Var}\,(V) = V^2 \left\{ -\frac{4}{N} + \frac{ad(a+d)+bc(b+c)}{(ad-bc)^2} \right.$$

$$\left. -\frac{3}{4}\left[\frac{(m-n)^2}{Nmn} + \frac{(r-s)^2}{(Nrs)}\right] + \frac{(ad-bc)(m-n)(r-s)}{2Nmnrs} \right\}$$

Suppose first that the association is positive or that ad > bc. Then C must fall on the line KL in the figure and

$$C = \frac{a - E(a)}{a(max) - E(a)}$$



To illustrate the derivation of the different formulas for cole's coefficient of interspecific association.

Let the species be so labeled that A is the less frequent species or $m \leq r$.

$$\text{Then } a(\max) = m \text{ and } C = \frac{a - mr/N}{m - mr/N}$$

$$= \frac{ad - bc}{ms}$$

Next suppose that there is negative association or that $ad<bc$. Then C must fall on the line JK and

$$C = \frac{a - E(a)}{E(a) - a(\min)}$$

The value of $a(\min)$ depends on whether $a \leq d$ or $a>d$. If $a \leq d$, $a(\min) = 0$. Writing C_1 for the coefficient in this case,

$$C_1 = \frac{a - mr/N}{mr/N}$$

$$= \frac{ad - bc}{mr}$$

If $a > d$, $a\ (min) = a-d$. Writing C_2 for the coefficient,

$$C_2 = \frac{a - mr/N}{mr/N - (a-d)} = \frac{ad - bc}{ns}$$

Cole also obtained the sampling variances of the three versions of his coefficient.

By the property

$$C = \begin{cases} +1 \text{ when } a = a\ (max) \\ -1 \text{ when } a = a\ (min) \end{cases}$$

Therefore, like Yule's Q, it suffers from the defect that no distinction is made between complete and absolute association.

2.1.4. THE SPACING OF THE QUADRATS:

When two species are found to be positively associated, the conclusion drawn is usually one or both of the following: (a) one of the species has a beneficial affect on the other, either directly or by modifying the environment in a way favorable to

it; or (b) some independent environmental factors are variable over the area, and because the two species have identical or overlapping tolerance ranges for the factors, both are forced to occupy coincident or overlapping areas.

The fact that a statistical test gives evidence that two species are positively associated does not, of course, lead directly to the conclusion that one of these mechanisms must be operating. The test by itself suggests only that the null hypothesis should be rejected. The null hypothesis is this: the probability that a quadrat contains species A is independent of whether it does or does not contain species B, and vice-versa. Rejection of the hypothesis, and the consequent acceptance of the alternative hypothesis, namely that the probabilities are not independent, does not automatically imply that it is the two species that are dependent. It may simply mean that the quadrats are dependent.

2.1.5. The effect of Quadrat Size

We turn now to a consideration of the effects of quadrat size on indication of association, assuming quadrat spacing to be

Clearly, only a limited range of size is permissible. The quadrats must not be so small that they are incapable of containing at least two individuals of the larger species. Nor must they be so large that one of the two species will occur in every quadrat, this is would cause one of the marginal totals of the 2 x 2 table to be zero and make a test impossible. For practical reasons the feasible range of quadrat size will often lie well within the theoretically permissible range.

2.2. Association Among K Species

Now we suppose there are K independent species. Examining a unit is equivalent to performing K independent Bernaulli trials with probabilities of success p_1 , p_2 , ... p_k respectively; the variate s takes the values 0,1 ... K. Since the outcomes of the trials are independent the pgf of s is

$$G(Z) = \prod_{j=1}^{k} g_j(Z) = \prod_{j=1}^{k} (q_j+p_jZ)$$

That is, G(Z) is the product of the pgf's of K independent bionomial distribution.

Let us write

$$H(Z) = (Q+P_z)^k$$

for the pgf of the approximating binomial. The mean and variance of the approximating distribution are therefore KP and KPQ respectively.

The mean and variance of the exact distribution are easily seen to be

$$E(s) = \sum_{j=1}^{k} p_j = KE(p) \qquad .....(I)$$

Where E(p) is the expectation of the p_j values; and

$$Var(s) = \sum_{j=1}^{k} p_j q_j = KE(p)\,[1-E(p)] - K\,Var(p) \qquad .....(II)$$

$$\text{Where Var } (p) = \left(\frac{1}{k}\right)\sum_{j}[p_j - E(p)]^2$$

is the variance of the p_j values.

The p_j's for j = 1, ... K are estimable from the data. An estimate of p_j is given by the observed proportion of units in which the jth species was found. Substituting the observed mean and variance of these estimates, say $\bar{p}$ and V(p), for their population values in (I) and (II) gives estimates of E(s) and Var(s). Then equating the latter estimates to the corresponding moments of the approximating (binomial) distribution of s,

We have $k\bar{p} = KP$ and $K\bar{p}(1-\bar{p}) - kv(p) = KPQ$.

Solving for P and K gives

$$P = \bar{p} + \frac{v(p)}{\bar{p}} \quad \text{and} \quad K = \frac{K}{1+v(p)/\bar{p}^{-2}}$$

as the desired parameters of the approximating binomial. One can now test the null hypothesis, that the species are mutually independent, by judging the fit of the approximating binomial distribution to the observed distribution of s.

2.3. Individuals in a Continuum

Continuum sampling is needed in the study of communities of sessile or sedentary organisms such as plants (on land or in fresh water) and benthic organisms (in salt or fresh water). Quadrats are the usual sampling units and in testing for association between two species, it is customary to treat each quadrat as if it were a discrete sample unit and to use the same methods as those described before.

Most ecologists (e.g., Greig -Smith, 1964) are aware of the problems that may arise from treating an arbitrary quadrat as though it were a discrete natural entity, but many authors seem to confound two wholly different sources of difficulty which ought to be treated separately. (a) the spacing of the quadrats and (b) the sizes of the quadrats.

To sample individuals that are scattered though a continuum usually entails taking arbitrary delimited bits of the continuum as sample units.

2.4. Segregation Between Two Species

When we examine the association between two species of plants the results will be strongly influenced by both the spacing of the quadrats and their sizes. This is because what is being investigated is not so much interspecies relationships per se but rather joint, two-species patterns, Other factors, besides the relationship between the species, affect these joint, patterns. This suggests that it would be worth while to attempt to study the pattern of each species in relation to the other without regard to the pattern of either in relation to the ground.

We assume that the plants occur as discrete, genetically distinct individuals, reproducing seed, and that therefore we shall not be misled by the presence of clumps of vegetative shoots which are, or may have been in the past, organically connected. What we now enquire is: do the two species form "relative clumps" ? A relative clump of species A, for instance, occurs in a group of plants in which the proportion of A's is greater than their proportion in the whole population. Likewise for species B. A relative clump may or may not be a spatial clump

also. Thus consider figure 2.4. In Figure 2.4(a) although the plants as a whole have a random spatial pattern, there is clear evidence of relative clumping. Conversely in figure 2.4b, although the population as a whole is strongly clumped, the two species are not clumped in relation to one another, since within each clump the A's and B's are present in the same proportions and are seldomly mingled.

The objective now is to study the relative patterns of two species independently of their spatial patterns. The first question that arises is: are the two species randomly mingled or are they relatively clumped ? If they are randomly mingled, they may be described as unsegregated; if not, they are to some extent segregated from each other.

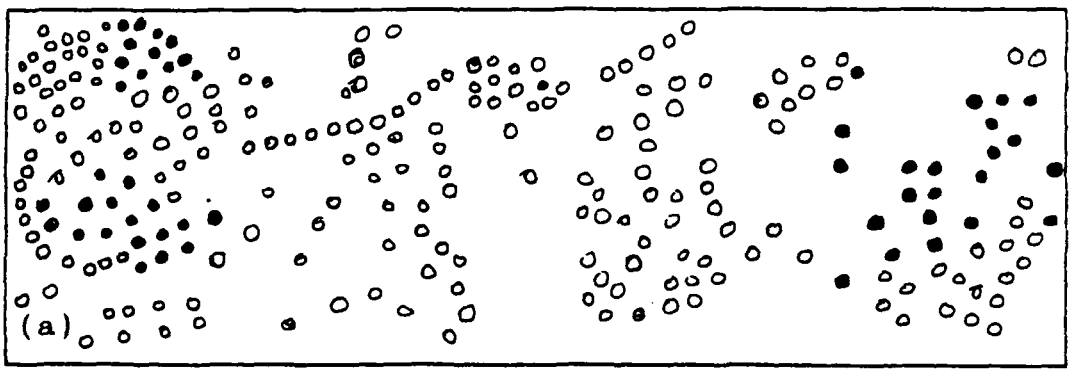


Figure 2.4a. A two species population in which all plants together have a random patterns, although the species are segregated.

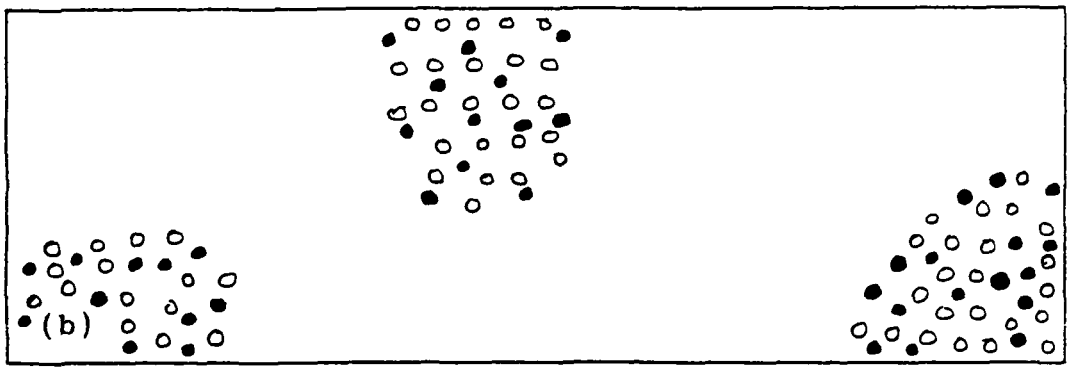


Figure 2.4(b). A two species population in which plants have clumped pattern but the species are unsegregated.

2.4.1. Segregation Among Many Species.

In the many-species, when the plants occur, not as distinct, discrete individuals, but as clumps or patches of appreciable area. The vegetation of swamps or of heaths or moors are examples. In drawing a map of such vegetation it would be impossible to represent individual plants by dimensionless dots, each marking a plant's center, as could be done with a map of a forest. In case of the vegetation it must necessarily be mapped as a many-phase mosaic, one phase of which might be bare ground.

CHAPTER III

SPECIES ABUNDANCE RELATIONS

CHAPTER III

SPECIES ABUNDANCE RELATIONS

3.0 Introduction:- Most ecological communities contain many species of organisms, and the species may vary greatly in their abundance from very common to very rare. Therefore as soon as one attempts to study whole communities rather than the interrelations among a few chosen species, the question immediately arises: how are the abundances of the different species distributed ? If there are N individuals belonging to 'S'species and the numbers of individuals in the respective species are N_1, N_2, ...N_S, have the N_j any consistent interrelationship, regardless of the type of community from which they come ? Attempt to answer this question have led to the development of "species-abundance" curves. If it should turn out that one single form of probability distribution with a small number of parameters (say two or three) fitted the data from the majority of observed communities, with only the parameter values varying from one community to another, interesting relationships might be discovered between the values of the parameters and the types of community they described.

In the present context "community" means all the organisms in a chosen area that belong to the taxonomic group the ecologist is studying. The chosen area is usually one that the ecologist regards as a convenient entity and is willing to consider as homogeneous in some intuitive sense. The reliance on intuition is necessary, since homogeneity can not be precisely defined at present; exactly what meaning, if any, should be attached to the term "homogeneous community" has for many years been hotly debated and no end to the discussion is in sight.

The same is true when it comes to defining the group of animals or plants that is constitute the community. To take all the living things in the specified area will not do. It would be impracticable to consider every kind of living thing in, say an acre of forest - the mammals, birds reptiles, amphibians, arthropods, and soil microfauna, together with the trees, shrubs herbs, ferns, mosses, and bacteria. A taxonomic group that the ecologist regards as an entity is usually chosen; often it is an entity only in the sense that it is a family, order or class (or other Taxon) that taxonomists are familiar with so that

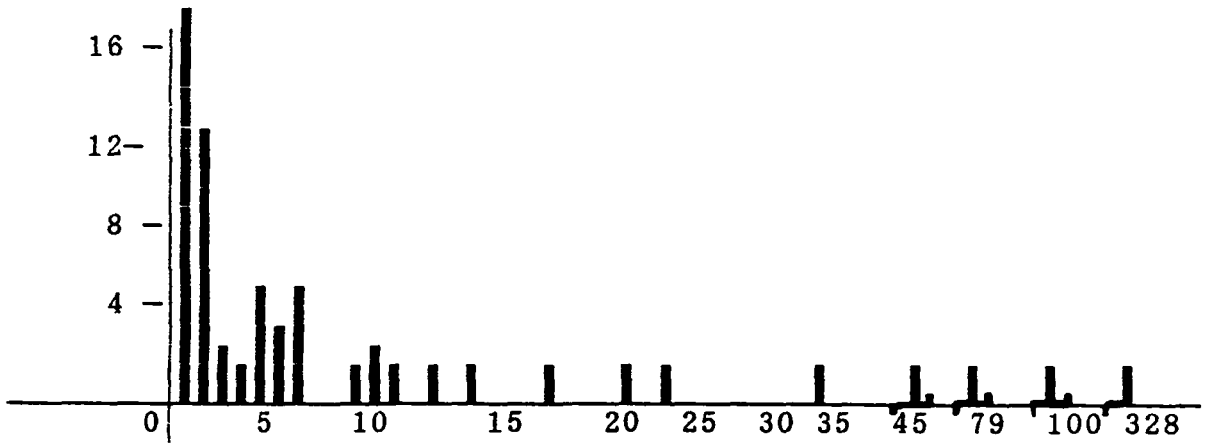
individuals can be fairly easily identified to species. The members of such a taxon that occur together at one place are designated a "taxocene".

Without attempting either to justify or to disparage the types of collection examined in attempts to determine their species- abundance relations, we now proceed with the mathematical theory that has been developed to account for them. In many collections it is found that singleton species (those represented by one individual) are numerous, often the most numerous. Species with successively more representatives, doubletons with 2, trebleton with 3, ... and so on, are usually progressively less numerous. Roughly speaking, one often finds many rare species and a few abundant ones, although, of course, in terms of numbers of individuals those of the few common species for out number those of the many rare species.

This frequently observed phenomenon has led to the method of tabulating species - abundance data customarily used: instead of listing the numbers of individuals in species 1, species 2, etc., we list the number of species, n_i, represented

by one member, ..., the number of species, n_r, represented by r members, ..., and so on. The n_r are, in fact, frequencies of frequencies. Next figure is an example.

3.1. Compound Poisson:- In all cases the collection at hand is treated as a random sample from some indefinitely large parent population. Assume further that each species is randomly dispersed; that is, the number of members the collection contains of, say, the jth species is a Poisson Variate with parameter λ_j. Then.



Number of species with 1,2,3, ... individuals in a collection of 822 individuals (52 species) of insects and mites. The individuals were adults emerging from a collection of fruiting bodies of the bracket fungus Foams fomentarias (Data from Pielore and Matheneman, 1966).

P_r (the jth species is represented by r members) = $\frac{e^{-\lambda}\lambda^r}{r!}$

Now consider all the species in the community. Their densities vary from species to species over a wide range. If there are S^* species in the whole population, we may regard the several values of λ as constituting a sample of size S^* from some continuous distribution of λ values having pdff(λ) Then the probability that any species will be represented by r members is

$$P_r = \int_0^\infty \frac{\lambda^r e^{-\lambda}}{r!} f(\lambda)d\lambda \quad \text{for} \quad r=0,1,2, \ldots\ldots (3.1.1.)$$

that is, the distribution of the different species frequencies $n_0, n_1, n_2, \ldots\ldots$ where $n_r = S^* p_r$ is assumed to have the form of a compound poisson distribution.

The observed distribution is a truncated form of the theoretical distribution, the zero class is missing. We do not in general, know the value of S^*, the number of species in the whole population, for presumably some of them will be missing from the collection, which is only a sample of the population.

Suppose the observed number of species is S. Then $S^{*}-S= n_{o}$ is the number of species represented by zero members in the collection, which is to say unrepresented.

It is worth contrasting this situation with that obtaining when the spatial pattern of one species (e.g. of plant) is being investigated by quadrat sampling. In the latter case we count the numbers of individuals of the one species concerned in a know number of different quadrats located at different places; thus we can count the number of empty quadrats (those from which this species is absent) and so obtain an empirical value of n_{o}.In obtaining the empirical distribution of species abundances in a collection, on the other hand, we are examining only a single area (equivalent to one quadrat) and counting the numbers of members it contains of each of S^{*}different species; since S^{*} is unknown, so also is n_{o}.

We now consider those members of the family of compound poisson distributions that have been fitted to observed species - abundance data.

3.2: Log Series Distribution

Suppose the values of λ for the different species are assumed to have a Type III distribution, that is $f(\lambda)$ in (3.1.1.) is given by

$$f(\lambda) = \frac{P^{-k}\,\lambda^{k-1}\,e^{-\lambda/p}}{\sqrt{(k)}}\,,\ \lambda \geq 2 \qquad \text{------(3.2.1)}$$

with K, P>0

Then P_r is a negative binomial variate, or

$$P_r = \frac{\sqrt{(k)+r)}}{r!\sqrt{(k)}}\left[\frac{P}{1+P}\right]^r\left[\frac{1}{1+P}\right]^k \quad \text{for } r=0,1,2, \text{ ----(3.22)}$$

3.2.1. The Discrete Lognormal Distribution:

Consider (3.1.1.) again. We shall now let $f(\lambda)$ be the pdf of the lognormal distribution that is, we assume the λ-values are a sample of size S^* from a distribution having pdf.

$$f(\lambda) = \frac{1}{\lambda\,\sigma\sqrt{2\pi}} \exp.\left[-\frac{1}{2\sigma^2}\left(\log\frac{\lambda}{m}\right)^2\right] \ldots\ldots(3.2.1.1.)$$

Equivalently, $\log\lambda$ is assumed to be normally distribution with p.d.f.

$$\phi(\log\lambda) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left[\log\frac{\lambda}{m}\right]^2\right] \ldots\ldots(3.2.1.2)$$

Then $E(\log\lambda) = \log m$ and $\mathrm{Var}(\log\lambda) = \sigma^2$ Notice that logm is the median as well as the mean of $\log\lambda$. Therefore m is the median value of λ, or the median abundance.

3.3. Negative Binomial:

$$f(\lambda) = \frac{1}{\Gamma(k)}\, p^{-k}\, \lambda^{k-1}\, e^{-\lambda/p} \qquad \ldots\ldots\ldots(3.3.1)$$

When the negative binomial distribution, with K>0, fits the data, it is possible to estimate S^* from observations on the sample. The probability that a species will contain r individuals is

$$p_r' = \frac{\Gamma(k+r)p^r}{r!\,\Gamma(k)(1+P)^{k+r}\,[1-(1+P)^{-k}]} \quad r = 1,2,\ldots \quad (3.3.2)$$

The mean and Variance of this distribution are

$$E(r) = \frac{KP}{1-(1+P)^{-k}}$$

and $Var(r) = (1+P+KP)\ E(r) - [E(r)]^2$

The parameters P and K may therefore be estimated from the mean and variance of the empirical distribution

Then since

$$S^* = \frac{S}{1-(1+P)^{-k}} ,$$

We may estimate S^* from S and the estimates of P and K. The sampling variance of the estimate is unknown.

3.3.1. The Geometric Distribution:

One may put K=1 in the pdf of $f(\lambda)$ in (3.3.1)

we get $f(\lambda) = \frac{1}{P} e^{-\lambda/p}, \qquad \lambda \geq 0$ .....(3.3.1.1.)

$$\text{and } p'_r = \left(\frac{P}{1+P}\right)^{\lambda-1} \frac{1}{1+P}, \qquad r = 1,2,..... \quad (3.3.1.2)$$

3.4. Diversity and it's Measurement:

When the species-abundance frequencies in an actual collection are well fitted by one or another of the theoretical distributions already described, the parameters of the fitted distribution are obviously suitable as descriptive statistics. If the distribution is lognormal, the appropriate statistics are the estimates of S^*, the total number of species in the population, and σ^2, the variance of the lognormal curve. If the distribution is negative binomial with $K \neq 0$, the appropriate statistics are the estimates of S^* and K (the parameter P depends on sample size).

What is needed are descriptive statistics that can be used for any community, no matter what the form of its species-

abundance distribution and even when no theoretical series can be found to fit the data.

We begin by considering the properties of any collection, regardless of whether it is to be treated as a population in its own right or as a sample from some larger parent population. Two statistics are clearly needed to describe a collection of which the first and most obvious is S, the number of species it contains. Now suppose we are dealing with data consisting of a list of numbers of individuals, $N_1, N_2, \ldots N_s$ in each of the S species. If the data are portrayed in histogram form, S is the range of the data or the width of the histogram.

As a second statistic, to describe the shape of the histogram, we require some thing analogous to variance. If the N_j were frequencies of some discrete quantitative variate, variance as ordinary calculated would, of course, be the obvious statistic to use, but we are now considering an unordered qualitative variate; the individuals are classified according to the species to which they belong and there is no a priori reason for listing

them in any particular order. The shape of the histogram is therefore best described in terms of what be called its "evenness". Thus the distribution has maximum evenness if all the species abundances (the N_j) are equal; and the greater the disparities among the different species abundances, the smaller the evenness.

"Diversity" is sometimes used merely as a synonym "number of species" or a single statistic in which the number of species and the evenness are confounded. A collection is said to have high diversity if it has many species and their abundances are fairly even. Conversely, diversity is low when the species are few and their abundances uneven. It will be seen that since diversity depends on two independent properties of a collection ambiguity is inevitable; thus a collection with few species and high evenness could have the same diversity as another collection with many species and low evenness.

This difficulty did not arise when the notion of adiversity was first introduced by Williams (see Fisher, Corbet and Williams, 1943). On the assumption that most species -abundances

distribution would be well fitted by logarithmic series distributions, he proposed that the parameter α of that distribution be used as an index of diversity. This index can be applied only if the logarithmic series does indeed fit the species-abundance data, but to determine whether it does may be impossible if there are only a few species and each is represented by a different number of individuals. Thus α is suitable as an index of diversity only if the collection at hand has many species and even then only if its species abundances form a logarithmic series. Some other measure of diversity is needed.

3.4.1. The Information Measure of Diversity:

As a measure of the diversity of the population, we wish to find a function of the p_j, $H'(p_1, p_2, \ldots p_s)$, say that meets the following conditions:

1. For a given S the function takes its greatest value when $p_j = 1/S$ for all j. Denote this greatest value by L(S). Then,

$$L(S) = H'\left[\frac{1}{S}, \frac{1}{S}, \ldots \frac{1}{S}\right].$$

2. $H'(p_1, p_1, \ldots p_s, 0, \ldots 0) = H'(p_1, p_2, \ldots p_s)$.

3. Suppose the population is subjected to an additional separate classification process that divides it into t classes, $B_1, B_2, \ldots\ldots B_t$. Each individual belongs to exactly one B-class, and the probability that it will belong to class B_k is q_k with $\sum_{k=1}^{t} q_k = 1$ Then the double classification yields St different classes, $A_j B_k$ (j=1..S; K=1, ...t).

Having specified the three conditions that H' is to satisfy, we now show that the only function with these properties is

$$H'(p_1, p_2, \ldots, p_s) = -C \sum_j p_j \log p_j, \quad \ldots\ldots\ldots (3.4.1.1).$$

3.4.2. Censused Communities:-

The Shannon function

$$H' = -\sum p_j \log p_j$$

as used in information theory is strictly defined only for an infinite population; it measures the information content of a code as distinct from a particular message in the code (Goldman 1953).

For a particular message, containing N symbols of S different kinds with N_j of jth kind ($\sum N_j = N$), the analogous measure is Brillouin's (1962) function, defined as

$$H = -\frac{1}{N} \log \frac{N!}{N_1 ! \, N_2 ! \, \ldots \, N_s !} \qquad \ldots\ldots\ldots (3.4.2.1.)$$

Margalef (1958) was the first to use this function to measure ecological diversity.

1. As Min $(N_j) \longrightarrow \infty$, $H \longrightarrow H'$

Where H is a measure of diversity in censused communities.

3.4.3. Sampled Communities:

If the diversity of a large community is to be estimated from a sample, and if we now write N_j for the number of

individuals of the jth species in the sample ($j=1,\ldots S; \sum N_j = N$), then

$$\hat{H}' = -\sum \frac{N_j}{N} \log \frac{N_j}{N} \qquad \ldots\ldots\ldots (3.4.3.1.)$$

is the maximum likelihood estimator of H'.

3.4.4. Hierarchical and Habitat Components of Diversity:

The classification of the individuals into species will be called the S-classification. There are S_i species in the ith genus, and N_{ij} individuals in the jth species of the ith genus

$$(j=1, \ldots . s_i ; \sum_{j=1}^{s_i} N_{ij} = N_i)$$

Now put:

H(G) for the genus diversity of community;

H(GS) for the species diversity of the community, that is, the "total" diversity;

$H_i(S)$ for the species diversity within the ith genus, and

ith genus, and

$$H_G(S) = \sum_{i=1}^{g} -\frac{N_i}{N} H_i(S)$$

for weighted mean of the species diversity in all g genera clearly,

$$H(GS) = \frac{1}{N} \log \frac{N!}{\prod_{i=1}^{g} \prod_{j=1}^{S_i} N_{ij}!}$$

$$H(GS) = H(G) + H_G(S) \qquad \ldots\ldots(3.4.4.1.)$$

Where $H_G(S)$, diversity within a genus average over all genera.

Let M_{ij} $(i=1,\ldots r;\ j=1,\ \ldots c)$ be the number of individuals of the ith species found in the jth habitat. Also, let $\sum_j M_{ij} = M_{i.}$ (the ith row total) be the total number of members of the ith species in

the community; in all habitats; and let $\sum_i M_{ij} = M_{\cdot j}$ (the jth column) be the total number of community members, of all species, found in the jth habitats. let M individuals comprising the community concerned have been fully censused so that the Brillouin index is the appropriate measure of diversity.

Now let the rowwise classification (by species) be called the A-classification and the columnwise classification (by habitats) be called the B-classification

Clearly

$$H(AB) = \frac{1}{M} \log \frac{M!}{\prod_i \prod_j M_{ij}!} \quad \ldots\ldots (3.4.4.2)$$

$$H(AB) = H(A) + HA(B)$$

3.4.5. The Measurement of Evenness:

In a fully censused community put N=SX+r, Where X=[N/S] and put Y = X+1 so that N=(s-r)X+rY

$$H(Max) = \frac{1}{N} \log \frac{N!}{(X!)^{S-r}(Y!)^{r}}$$

$$H(Min) = \frac{1}{N} \log \frac{N!}{(1!)^{S-1}(N-S+1)!}$$

A convenient measure of evenness is now given

(Hurlbert, 1971) by $$V = \frac{H - H(min)}{H(max) - H(min)} \quad \ldots\ldots(3.4.5.1)$$

For large community

$$H'(min) = \lim_{n \to \infty} H(min) = 0$$

$$H'(max) = -\sum \frac{1}{S^*} \log \frac{1}{S^*}$$

$$= \log S^*$$

$$V' = \frac{H' - H'(min)}{H'(max) - H'(min)} = \frac{H'}{\log S^*}$$

with $$\tilde{V}' = \frac{\tilde{H}'}{\log S^*} \quad \ldots\ldots\ldots(3.4.5.2.)$$

$$\text{Var}\,(\tilde{V})' = \frac{\text{Var}(\tilde{H}')}{(\log S^*)^2} \qquad \ldots\ldots\ldots\ldots(3.4.5.3)$$

3.4.6. Simpson's Measure of Diversity

Renyi's Entropy of ordex α

$$H_\alpha = \frac{\log \sum_{i=1}^{S} p_i^{\alpha}}{1 - \alpha}$$

Where $\lim_{\alpha \to 1} H_\alpha = H'$ Shannon's Entropy.

Another measure of information as suggested by

Simpson $H_2 = -\log \sum p_i^2 \qquad \ldots\ldots\ldots(3.4.6.1.)$

Which can be get putting $\alpha=2$ in Renyi entropy. The function $\sum p_i^2 = \lambda$, is the probability that any two individuals picked independently and at random from the community will belong to the same species. Where λ as a measure of concentration or dominance and "expected commonness".

If λ is high, we can take $1-\lambda$ is a measure of diversity.

CHAPTER IV

DISCOVERING A NEW SPECIES

CHAPTER IV

DISCOVERING A NEW SPECIES

4.0 Introduction:

Population consists of a number of unknown species, possibly countably many. The number of species in a population provides a partial description of the population and may be used in the comparison of population over time or space.

Finding number of species in a population is a problem, because of the construction and nature of the population. Many species are very rare and there is every possibly of missing so many of them in the sample. And it is not only rarity of the species their habitat (nature of living) is also a point that we may miss some of them in the sample.

4.1. Estimating the Probability:

We search the population by selecting one member at a time, noting its species identity and returning it to the population. A search is called an n-stage search if n selections are made.

Imagine that the species are labeled 1,2, ... in any arbitrary fashion. Let p_i denote the probability that a randomly selected members belongs to the ith species, i=1,2, ... and let X_i^n be the numbers of representatives of the species i in the n stage search. As indicated in Starr (1979), the conditional probability that we will discover a new species in the n+1st selection given the X_i^n is

$$Un = \sum_i p_i \; I\left[X_i^n = 0\right]$$

The unconditional probability that at the last stage of an n+1 stage search we will find a new species is equal to

$$\theta_n = EU_n = \sum_i p_i \; (1-p_i)^n.$$

We are particularly interested in finding estimators of θ_n, which are to be used as predictors of U_n. An estimator obtained by extending the initial search an additional stage has been discussed extensively in a number of previous papers, including Starr (1979) and Robbins (1968). Based on the search of size n+1,

the estimator is

$$V_1 = q_1(n+1)/(n+1)$$

$$\text{Where } q_k(n+1) \sum_i I\left[X_i^{n+1} = k\right]$$

denotes the number of species which have k representatives in the n+1 stage search, $k \geq 1$.

Robbins (1968) has shown that V_1 is a good predictor of U_n in the sense that

$$EV_1 = EU_n = \theta_n$$

$$\text{and } E(V_1 - U_n)^2 < (n+1)^{-1}$$

Starr (1979) generalized V_1 to a class of estimators which he called Robbins type estimators. In his study, the original search was extended by an additional m stages, $m \geq 1$. Starr show that

$$V_m = \sum_{k=1}^{m} \frac{\binom{m-1}{k-1}}{\binom{n+m}{k}} q_k(n+m)$$

is the unique linear combination of

$$q_k\ (n+m) = \sum_i I\left[X_i^{n+m} = k\right], \qquad k = 1, \ldots. \ n+m,$$

Which has expectation θn.

Starr (1979, page 650) conjectured that V_m is the uniformly minimum Variance unbiased estimator (UMVUE) of θ_n. In section 4.1.2 we shall disprove Starr's conjucture by obtaining the UMVUE of θ_n for a special case. The UMVUE of θ_n is further compared in section 4.1.3 with Robbin's estimator in the associated prediction problem.

4.1.2. Results: Assume that the initial search has been extended an additional m stages, $m \geq 1$. Let

$$(1)\ d = d(n+m) = \sum_{k=1}^{n+m} q_k\ (n+m)$$

represent the number of observed species in the search of size n+m.

Theorem 1. Suppose there are μ species, $\mu \leq n+m$ and $p_1 = p_2 = \ldots p_\mu = \mu^{-1}$. Then the UMVUE of θ_n based on a search of size n+m is

$$W_m = W_m(d) = \sum_{k=0}^{m-1} \binom{m-1}{k} \alpha_{d-1,n+k} \Big/ \alpha_{d,n+m}$$

Where $\alpha_{p,q}$ are the Stirling numbers of the second kind, $p \leq q$, defined by $x^q = \sum_{p=1}^{q} \alpha_{p,q}\, x^{(p)}$; if $p > q$, we define $\alpha_{p,q} = 0$.

Proof:- Under the specified condition, Harris (1968) showed that d in (1) is the complete sufficient statistic for μ. Thus, from the Lehmann-Scheffe' theorem, we can conclude after some manipulations that the unique UMVUE $W_m(d)$ of θ_n is given by $Wm(d) = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \alpha_{d,k+m} \Big/ \alpha_{d,n+m}$.

The result follows directly from the following identity

$$\sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \alpha_{d,k+m} = \sum_{k=0}^{m-1} \binom{m-1}{k} \alpha_{d-1,n+k}$$

See Chao (1980) for further details.

Remarks:- (A) We can show that, based on the original search, no unbiased estimator which is a function of the complete sufficient statistic exists. Similarly no, such estimator can be obtained from a search of size less than n.

(B) If m=1, the UMVUE of θ_n can be written as

$$W_1 = 1-d/\left[\frac{\alpha_{d,n+1}}{\alpha_{d,n}}\right]$$

with d=d(n+1) defined by (1). It is interesting to find that W_1 is analogous in form to U_n for the uniform case, since $U_n=1-d(n)/\mu$, where d(n) is the number of observed species in the initial search. Actually, according to a result provided in Harris (1968) for a sample of size n+1, $\alpha_{d,n+1}/\alpha_{d,n}$ is asymptotically the UMVUE of μ.

We now proceed to examine the asymptotic behavior of W_m.

Theorem 2. If $n \longrightarrow \infty$ and $\mu \longrightarrow \infty$ in such away that $n/\mu, \longrightarrow \alpha$ $0<\alpha<\infty$ then with probability one,

$$W_m = \exp(-R_m) + O(n^{-1})$$

Where R_m is the unique solution of

$$f(R) = \frac{R}{\{1-\exp(-R)\}} = \frac{(n+m)}{d}, \qquad m = 1, 2 \ldots\ldots$$

Proof:- It follows from the recursive formula for the Stirling numbers of the second kind (Jordan 1950, page 169) that

$$W_m = \sum_{k=0}^{m-1} \begin{bmatrix} m-1 \\ k \end{bmatrix} \left(\alpha_{d,n+k+1} - d\alpha_{d,n+k} \right) / \alpha_{d,n+m}$$

$$= \left| - \frac{d\alpha_{d,n+m-1}}{\alpha_{d,n+m}} \right| + Q,$$

Where

$$Q = \sum_{k=1}^{m-1} \binom{m-1}{k-1} \frac{\alpha_{d,n+k}}{\alpha_{d,n+m}} \qquad \sum_{k=0}^{m-2} \binom{m-1}{k} \frac{d\alpha_{d,n+k}}{\alpha_{d,n+m}}$$

Applying the results

$$\frac{\alpha_{d,n+m-1}}{\alpha_{d,n+m}} = \frac{R_m}{n+m} \{1+0(n^{-1})\}$$

$$\frac{d\alpha_{d,n+m-1}}{\alpha_{d,n+m}} = \{1-\exp(-R_m)\}\{1+0(n^{-1})\},$$

Which are deduced from a theorem of Harris (1968, page 841), we can established that $Q=O(n^{-1})$. The result is immediate.

Starr (1979) found that the Robbins predictor V_1 has an unattractive property, namely V_1 is strongly negatively for related with U_n ↗ If we employ $\exp(-R_m)$ as a predictor of U_n, a similar drawback excists. The negative correlation can be explained in the following intuitive way: the more species we found in the search, the more likely we are to

discover a new species in the next selection. Note that the negative correlations are asymptotic results. Therefore the we are essentially assuming that there are many species, so that negative correlations can be reasonably understood. Numerical results indicate that W_m increases from 0 to 1 as d is increased from 1 to n+m for any μ. Then the negative correlation is still valid even if there are few species.

4.1.3. Comparison:

We now compare the performance of W_1 as a predictor of U_n with that of V_1. It will be shown that W_1 is the better predictor in the sense that $E(W_1-U_n)^2$ is asymptotically uniformly smaller than $E(V_1-U_n)^2$. Robbins (1968) showed that

$$(n+1)\ E(V_1-U_n)^2 \longrightarrow f_1(\alpha) = e^{-\alpha}\ (1+\alpha)-e^{-2\alpha},$$

If $n,\mu \longrightarrow \infty$ such that $n/\mu \longrightarrow \alpha$, $0 < \alpha < \infty$. Under the same conditions, we can establish that

(2) $(n+1)\ E(W_1-U_n)^2 \text{ -----} > f_2(\alpha)$

Where $f_2(\alpha)=\alpha^2 e^{-2\alpha}(1-e^{-\alpha}-\alpha e^{-\alpha}) +\alpha e^{-\alpha}(1-e^{-\alpha}-\alpha e^{-\alpha})+2\alpha^2 e^{-2\alpha}$

The proof of (2) is omitted, although the derivation is indirect. The reader is referred to Chao (1980, pages 13-16) for details.

We show that the UMVUE is also superior in the associated prediction problem by claiming that
$f_2(\alpha) < f_1(\alpha)$ for all $0<\alpha<\infty$. It is equivalent to verify that

$$g(\alpha) = e^{-\alpha}(2+\alpha^2-e^{-\alpha}) <1,$$

Which follows from the fact that $g(\alpha)$ is a strictly decreasing function on $[0,\alpha]$ and consequently $g(\alpha) < g(0) = 1$, for all $0 < \alpha < \infty$.

Although $E(W_1-U_n)^2$ is uniformly smaller than $E(V_1-U_n)^2$, we still do not know whether $E(W_1-U_n)^2$ will attain the minimum in the class of all unbiased estimators of θ_n. We finally remark

that $E(V_1-\theta_n)^2 - E(W_1-\theta_n)^2$ is asymptotically equal to $E(V_1-U_n)^2 - E(W_1-U_n)^2$. This fact reveals that the difference in variance when we employ the UMVUE, instead of V_1, to estimate θ_n is essentially the reduction of mean square error if W_1, rather than V_1, is used to predict U_n.

4.2. Linear Estimation:

We consider a population π composed of (possibly countably many) distinct species, which we imagine to be labelled with the integers 1,2, ... in some arbitrary fashion. Let p_i denote the probability that an object chosen from π is a representative of species i; we suppose that there is no a priore information available concerning either the number of species in the population or the vector of search probabilities $p=(p_1,p_2,..)$ except the $P \in S$, where

$$S = \left\{\left[p_1 p_2,.....\right): 0 \leq P_i \leq 1 \ \forall_i \text{ and } \Sigma_i p_i = 1\right\}$$

We may search the population by selecting one member of π at a time, noting the species to which it belongs, and returning it to the population. (If $p_i>0\ \forall_i$ and π is infinite, then an equivalent search may proceed nonsequentially without replacement). If n independent selections are made then we say the search has size n (or is n-stage), let X_i^n denote the random number of representatives of species 'i' that will be found in the search, i=1,2, ..., and say that species i has been discovered if X_i^n assumes a positive value.

The quantity of interest in this note is the realization of the unobservable random variable.

$$U_n = \sum_i p_i I\left[X_i^n = 0\right]$$

the sum of the unknown probabilities associated with species which will not be discovered in a search of size n. U_n may be regarded as the random conditional probability that we will discover a new species at the last stage of an n+1 stage search, that is, given the values $X_i^n = x_i^n$, i=1,2 resulting from a

search of size n the realization.

$$u_n = \sum_i p_i \; I\left[x_i^n = 0\right]$$

of U_n is the conditional probability that if the search were extended one more stage we would discover a new species.

In this note we discuss the problem of estimating U_n. Bear in mind that the available data comprise only sample frequencies for those species which have been discovered and that the labelling is that of the searcher. To put this perspective. suppose that at the conclusion of a search of size n a total of d species have been discovered, and that their frequencies are $X_i^n = x_i^n$, $i=1, \ldots d$, Where the indices are imposed by the searcher in some arbitrary manner, for example, the order in which the species were discovered. Then our problem may be formulated in the following way.

Imagine that there are a total of d+1 species in the population with search probabilities $(p_1, \ldots p_d, U_n)$ and

with corresponding sample frequencies $(x_1^n, \ldots x_d^n\ 0)$. How many we use this data set to estimate Un ? Two observations are immediate. If we view the data from this perspective (that is, conditionally), then knowledge of the number of species in the population, say K, is irrelevant to estimation unless d=k in which case we know certainly that $U_n=0$. Moreover, it is apparent that standard procedures, such as maximum liklihood (which estimates U_n to be zero for every n), are in adequate.

However, an indirect method has surfaced in the literature, apparently suggested by A.M. Turning (see[2]) and discussed in a variety of detail and perspective by Good [2,3], Good and Toulmin [4], Harris [5], Knott [6], Robbins [7], and their bibliographies. Consider the quantity

$$\theta_n = E(U_n) = \sum_i p_i\ EI\left[X_i^n = 0\right] = \sum_i p_i q_i^n$$

Where we have set $q_i = 1-p_i$, $\forall_i$. θ_n denotes the unconditional probability that at the last stage of an n+1 stage search we will

discover a new species. To see this directly let A be the event that a speceis will be discovered at stage n+1, and let A_i denote the event that species i will be discovered at stage n+1; then the A_i are mutually exclusive with geometric probability $p_i q_i$ for each i, and $A = U_i A_i$, so that

$$P(A) = \sum_i P(A_i) = \theta_n .$$

Suppose now that we can develop an estimator V of θ_n for which $E(V) = \theta_n$. Then, since $E(U_n) = \theta_n$, there is some reason to hope that realizations of both V and U_n will be close to θ_n with high frequency, and hence close to one another, so that a given realization of V may represent a useful estimate of U_n. Thus common to the papers cited above is the attempt to develop and study estimators of θ_n. Unfortunately, the problem of judging the goodness of such estimators when they are utilized to predict U_n. apperars to have received only modest attention, and that from a single perspective.

Of special interest to us here are estimators of the type proposed by Herbert Robbins[7]. Suppose that an initial search of size n is completed, at which time the random variable U_n assumes the unobservable value U_n. Assume, however, that with the objective of improving our chances of accurately predicting U_n, we extend the search one additional stage, and let

$$q_k\ (n+1) = \sum_i I\left[X_i^{n+1} = K\right]$$

denote the number of species with exactly k representatives in the extended search of total size n+1. Robbins proposed as a predictor (he regards it as an "estimator" of U_n the random Variable

$$V_1 = \frac{q_1(n+1)}{n+1},$$

the proportion of species with exactly one representative in the extended search. V_1 represents a good predictor of U_n in the sense that

(1) $E(V_i) = E(U_n) = \theta_n$ and $E(V_i - U_n)^2 < \frac{1}{n+1}$

for every $p \in S$ (see [7]).

Indeed, we shall prove that V_i is the unique linear combination of $q_1(n+1)$, $q_2(n+1)$, $\ldots q_{n+1}(n+1)$ with expectation θ_n. However, V_i does not follow U_n in a sense that might reasonably be demanded of a predictor; viz that realizations u_n of U_n larger than θ_n be accompanied with high frequency by realizations v_i of V_i larger than θ_n, and vice versa. In particular we shall prove that if the search probabilities pi are equal and if the size of the search is of the same order as the number of species, then V_i and U_n are strongly negatively correlated. We conclude that althogh V_i will be close to U_n in the average sense of (1), that this is largely a result of the fact that the random variables have a common mean and modest variances, rather than a consequence of their being positively related or associated in any commonly understood sense of predictive inference. θ_n the other hand, we hasten to observe that we do not yet know how to

do better (and perhaps cannot).

In the next section we shall consider a class of predictors of U_n obtained by extending an initial search of size n by an additional m stages, and call it the class of Robbins type predictors (Robbins studied the case m=1). One of the referees has envisioned the following kind of conversation that could result at the conlcúsion of an n-stage search from the use of Robbins-type prediction. Paraphrased it goes:

Searcher: "I am considering making one more search. If I do so, am I likely to discover a new species ?"

Statistician: "Make the search and then I will tell you". The reference becomes less awkward if the problem is developed in a design context; for example:

Searcher: "I am contemplating extending my initial search an additional large number M of stages, and will so do if the

expected number M. u_n of inviduals I will select in the second search who do not represent species discovered in my initial search is large. Who do you recommend"?

Statistician: "Make one more search and then I will tell you".

4.2.1. Results: Suppose that an initial search of size n has been extended an additional m stages, m=1,2,... and let

$$q_k\ (n+m) = \sum_i I\left[X_i^{n+m} = K\right]$$

denote the number of species for which there will be exactly K representatives in the search of total size n+m. Our immediate objective is to use the values q_k (n+m), K=1,2,.... n+m, to estimate the parametric function.

$$\theta_n = \sum_i p_i q_i^n.$$

Theorem: 1: Let $\alpha_o, \alpha_1, \ldots \alpha_{n+m}$ be constants and define

$$W_m = \alpha_o + \sum_{k=1}^{n+m} \alpha_k q\ (n+m)$$

Then $E(W_m) = \theta_n$ identically in $P \in S$ if and only

if $\alpha_o = \alpha_{m+1} = \ldots\ldots\ldots = \alpha_{n+m} = 0$ and

$$\alpha_k = \frac{\binom{m-1}{k-1}}{\binom{n+m}{k}} \quad \text{for } K=1, \ldots.m,$$

وبذلك

$$V_m = \sum_{k=1}^{m} \frac{\binom{m-1}{k-1}}{\binom{n+m}{k}} q_K(n+m)$$

is the unique linear from in $\left\{q_k(n+m),\ k=1,\ \ldots.(n+m)\right\}$

with expectation θ_n.

Proof:- Observe that the random variables

$\left\{q_k\ (n+m),\ k=1,\ \ldots..\ n+m\right\}$ are constrained by the condition.

(2) $\sum_{k=1}^{n+m} kq_{iK}(n+m) = n+m.$

By direct computation

$$E\ W_m = \alpha_o + \sum_{k=1}^{n+m} \alpha_k \sum_i E\left\{ I\left[X_i^{n+m} = K \right] \right\}$$

$$= \alpha_o + \sum_{k=1}^{n+m} \alpha_k \binom{n+m}{k} \sum_i p_i^k\ q_i^{n+m-k}$$

$$= \alpha_o + \sum_{k=1}^{n+m} \alpha_k \binom{n+m}{k} \sum_i p_i \sum_{j=0}^{k-1} (-1)^j \binom{k-1}{j} q_i^{n+m-k+j}$$

Interchanging the order of summation and making use of symmetry in the arguments of the binomial coefficients yield.

$$(3)\ E\ W_m = \alpha_o + \sum_i p_i \sum_{j=1}^{n+m} \sum_{k=0}^{n+m-j} (-1)^k \alpha_{k+j} \binom{n+m}{k+j} \binom{k+j-1}{k} q_i^{n+m-j}.$$

Too see that $E\ V_m = \theta_n$, set $\alpha_o = \alpha_{m+1} = \ldots\ldots\ldots \alpha_{m+n} = 0$ and

$$\alpha_k = \frac{\binom{m-1}{k-1}}{\binom{n+m}{k}},\ k=1,\ \ldots\ldots,m$$

in (3). Then

$$E\ V_m = \Sigma_i\ p_i \sum_{j=1}^{m} \sum_{k=0}^{m-j} (-1)^k \binom{m-1}{k+j-1} \binom{k+j-1}{k} q_i^{n+m-j}$$

$$= \Sigma_i\ p_i\ q_i^n + \Sigma_i\ p_i \sum_{j=1}^{m-1} \sum_{k=0}^{m-j} (-1)^k \binom{m-1}{k+j-1} \binom{k+j-1}{k} q_i^{n+m-j}$$

$$E\ Vm = \Sigma_i\ p_i\ q_i^n + \Sigma_i\ p_i \sum_{j=1}^{m-1} \binom{m-1}{j-1} q_i^{n+m-j} \sum_{k=0}^{m-j} (-1)^k \binom{m-j}{k} = \theta_n$$

Where the last equality follows from the Binomial theorem.

To prove uniqueness, set

$$a_j = \sum_{k=0}^{n+m-j} \left[-1\right]^k \alpha_{k+j} \binom{n+m}{k+j} \binom{k+j-1}{k}$$

for each j=1, ... n+m-1. Then from (3) we have for every p∈S.

$$(4)\ E\ (W_m) = \alpha_o + \alpha_{n+m} + \sum_{j=1}^{n+m-1} a_j\ \Sigma_i\ p_i\ q_i^{n+m-j}$$

Thus, setting $b_j = a_j$ for $j \neq m$ and $b_m = (a_m - 1)$,

it is easily seen that $E(W_m) = \theta_n$

identically in $p \in S$ only if

$$(5)\quad \alpha_o + \alpha_{n+m} \sum_{j=1}^{n+m-1} b_j \sum_i p_i q_i^{n+m-j} = 0$$

identically in $p \in S$. Clearly, (5) can hold only if $\alpha_{n+m} = -\alpha_o$, where α_o is arbitrary. Let b denote the column vector whose jth component is b_j, and for given $p \in S$ let $d_{(p)}$ denote the column vector with jth component $\sum_i p_i q_i^{n+m-j}$, $j = 1, \ldots n+m-1$. Thus from (5) we have that $E(W_m) = \theta_n$ identically in $p \in S$ only if

(6) $b' d(p) = 0$ for every $p \in S$.

In the sequel we shall show that

(7) span $\{d_{(p)}, p \in S\}$ has dimension n+m-1. So that (6) holds only if b is null vector; that is, only if

$a_j = 0$, $j = 1, \ldots n+m$, $j \neq m$ and $a_m = 1$.

To summarize, $E\, W_m = \theta_n$ identically in $p \in S$ only if

$$\sum_{k=0}^{n+m-j} (-1)^k \alpha_{k+j} \binom{n+m}{k+j} \binom{k+j-1}{k} = 0, j=1, \ldots\ldots n+m-1, j \neq m.$$

$$\sum_{k=0}^{n} (-1)^k \alpha_{k+m} \binom{n+m}{k+m} \binom{k+m-1}{k} = 1, \alpha_{n+m} = -\alpha_0$$

and α_0 is arbitrary. Thus if $E\, W_m = \theta_n$, any choice of α_0 uniquely determines the other coefficients $\alpha_1, \alpha_2, \ldots\ldots \alpha_{n+m}$. But for an arbitrary choice of a constant α it follows from (2) that

$$v_m = (1-\alpha)\, v_m + \alpha v_m$$

$$= (1-\alpha) \sum_{k=1}^{m} \frac{\binom{m-1}{k-1}}{\binom{n+m}{k}} + \alpha \sum_{k=1}^{m} \frac{\binom{m-1}{k-1}}{\binom{n+m}{k}} \left\{ \frac{1}{k} \left[(n+m) - \sum_{\substack{j=1 \\ j \neq k}}^{n+m} j q_j(n+m) \right] \right\}$$

$$= \alpha_0 + \sum_{k=1}^{n+m} \frac{\alpha}{k}\, q_k(n+m),$$

Where $\alpha_o = \alpha(n+m)\sum_{k=1}^{m} \binom{m-1}{k-1}/k\binom{n+m}{k}$ is arbitrary (since α is) and $\alpha_1, \ldots\ldots \alpha_{n+m}$ are uniquely determined by the choice of α_o. Thus, $W_m = V_m$ if $E\, W_m = \theta_n$ are uniquely determined by the choice of α_o. Thus .

$$W_m = V_m \text{ if } E\, W_m = \theta_n.$$

It remains to verify (7). For each j=1, ... n+m-1 consider the column vector C_j whose kth component K=1, ... n+m-1, is the kth component of $d_{(p)}$, where p is the k+1 component vector p = (1/(k+1)), 1/(k+1), ... 1/(k+1), so that p∈S. From the (n+m-1)x (n+m-1)matrix C with jth column C_j, so that

$$C = \begin{bmatrix} \left(-\frac{1}{2}\right)^{n+m-1} & \left(\frac{1}{2}\right)^{n+m-2} \cdots\cdots\cdots \left(-\frac{1}{2}\right) \\ \left(\frac{2}{3}\right)^{n+m-1} & \left(\frac{2}{3}\right)^{n+m-2} \cdots\cdots\cdots \left(\frac{2}{3}\right) \\ \left(1-\frac{1}{n+m}\right)^{n+m-1} & \left(1-\frac{1}{n+m}\right)^{n+m-2} \cdots \left(1-\frac{1}{n+m}\right) \end{bmatrix}$$

Then the determinent of C is easily seen to be propertional to the Vandermonde determinant which is nonzero, establishing (7), and completing the proof.

Remarks: 1.

It follows from (5) that for $m \leq 0$, there is no choice of α_o, ... α_{n+m} for which $E(W_m) = \theta_n$ identically in $p \in S$; that is, no linear form in $\{q_k(n+m),\ k=1,\ \ldots.\ n+m,\}$ obtained from a search of size less than n+1 has expectation θ_n for every $p \in S$. This contradicts the assertion (2.09) of Knott [6] that

$$\sum_{i=1}^{n} \frac{(-1)^{i} q_i(n)}{\binom{n}{i}}$$

is an unbiased estimate of θ_n. In particular the estimator

$$(8) \qquad Vo = \frac{q_1(n)}{n}$$

of Good [2] has bias $E(Vo-\theta_n) = \sum_i \left(\frac{p_i}{q_i}\right) p_i\, q_i^n$

To verify the remark, take p to be the K component vector (1/k, ... 1/k); then from (5) for $m \geq 0$, $E(W_m)=\theta_n$ identically in $p \in S$. only if in particular

$$(9)\quad \alpha_o + \alpha_{n+m} \sum_{j=1}^{n+m-1} a_j \left(1- \frac{1}{k}\right)^{n+m-j} = \left(1- \frac{1}{k}\right)^{n}$$

for every k=2,3,... clearly there is no set of n+m+1 coefficients for which (9) holds identically in k, proving the remark.

Remark. 2: We suspect (but have not yet proved) that V_m is the uniformly minimum variance unbiased estimator of θ_n based on a search of size n+m.

Next, we turn our attention specifically to V_1, the predictor of U_n based on a search of total size n+1. The difficulty with V_1 referred to in the introduction is exhibited as

Theorem 2. suppose that there are k species, that $p_1 = \ldots\ldots p_k$, and that n and k become large in such a way that

$$(10) \quad \frac{n}{k} \longrightarrow \alpha \qquad 0 < \alpha < \infty$$

Then, under the limiting operation defined by (10)

$$(11) \rho(V_1, U_n) \longrightarrow f_\rho(\alpha) = -\frac{\alpha^2}{\left[\left(\alpha e^{\alpha} - \alpha^2 - \alpha\right)\left(e^{\alpha} - \alpha^2 + \alpha - 1\right)\right]^{1/2}}$$

Where ρ denotes correlation

Proof. For any $p \in S$.

$$E\, U_n = EV_1 = \sum p_i\, q_i^n$$

$$E\, U_n^2 = \sum p_i^2 q_i^n + \sum_{i \neq j} \sum p_i\, q_j (1 - p_i - p_j)^n$$

$$E\, V_1^2 = \frac{1}{(n+1)} \left[\sum {}_i p_i q_i^n + n \sum_{i \neq j} \sum p_i p_j (1 - p_i - p_j)^{n-1} \right]$$

$$E\ U_n\ V_i = \frac{1}{(n+1)} \left[\sum_i p_i q_i^n + \sum_{i \neq j} \sum p_i p_j\ (1-p_i-p_j)^n \right.$$

$$\left. \ldots \ldots \ldots \ldots + n \sum_{i \neq j} \sum p_i p_j q_i\ (1-p_i-p_j)^{n-1} \right]$$

Setting $p_i = 1/k$, $i=1, \ldots k$ and (carefully) taking the limit defined by (10) as n and k tend to infinity yield

$$(n+1)\ \mathrm{Cov}\ (V_i, U_n) \longrightarrow f_c(\alpha) = -\alpha^2 e^{-2\alpha}$$

$$(n+1)\ \mathrm{Var}\ (U_n) \longrightarrow f_u(\alpha) = \alpha e^{-\alpha} - \alpha^2 e^{-2\alpha} - \alpha e^{-2\alpha}$$

$$(n+1)\ \mathrm{Var}\ (V_i) \longrightarrow f_v(\alpha) = e^{-\alpha} - \alpha^2 e^{-2\alpha} - e^{-2\alpha} + \alpha e^{-2\alpha}$$

The result is immediate.

Remarks:-

(1) The limit $f_\rho(\alpha)$ of the correlation functions is increasing in α, tends to -1 as $\alpha \longrightarrow 0$ and to 0 as $\alpha \longrightarrow \infty$. The values of $f\rho$ are given in Table 1 for various α.

Table 1

Values of $f_\rho(\alpha)$, defined by (11) as a function of α

α	0.1	0.2	0.3	0.4	0.5
$f_\rho(\alpha)$	-0.9954	-0.9900	-0.9835	-0.9759	-0.9671
α	0.6	0.7	0.8	0.9	1.0
$f_\rho(\alpha)$	-0.9569	-0.9452	-0.9319	-0.9169	-0.9001
α	1.5	2.0	3.0	5.0	10.0
$f_\rho(\alpha)$	-0.7896	-0.6444	-0.3582	-0.0830	-0.0014

2. The limit f_u of the variance of U_n is maximized at the α value which solves $e^{\alpha}(1-\alpha) = 1-2\alpha^2$, and the limit f_v of the variance of V_1 at the α value which solves $3-4\alpha + 2\alpha^2 = e^{\alpha}$. Thus, the maximum values of f_u and f_v, achieved at about $\alpha = 1.97$ and 0.46 respectively, are approximately 0.16 and 0.33; f_u and f_v zero tend to as α ——> zero or infinity.

3. From (12) it follows that

$$(n+1)\, E(V_1 - U_0)^2 \longrightarrow e^{-\alpha}(1+\alpha) - e^{-2\alpha}$$

agreeing with [7]. The limiting quantity has a maximum value of about 0.61.

4. We do not know whether Robbins -type predictors of U_n may be positively correlated with U_n for some choice of m>2. However, for m =2, $\lim \rho(V_2, U_n) = \lim \rho(V_1, U_n)$ and $\lim \mathrm{Var}(V_2) = \lim \mathrm{Var}(V_1)$.

5. Concerning the predictor V_o of U_n defined by (8) we have also that $\lim \rho(V_o, U_n) = \lim \rho(V_1, U_n)$ and $\lim \mathrm{Var}(V_o) = \lim \mathrm{Var}(V_1)$ [See(1)].

4.3. Nonparametric Estimation:

We take sampling sequentially, and it is in this context that a related quantity arises: the probability of discovering a new species in a future sample based on sampling that has already taken place. By it self, this probability indirectly leads to information about the number of species in the population; it might also be used in a sequential sampling scheme where the goal is to decide when to stop sampling.

We know that the conditional probability of discovering a new species in one additional search is

$$U_n = \sum_i p_i \; I(X_i^n = 0), \qquad .....(4.3.1)$$

Where $p_i = Pr(X_i = i)$. The corresponding unconditional probability of new species discovering is

$$\theta_n = E(U_n) = \sum_i p_i q_i^n \qquad .....(4.3.2)$$

Where $q_i = 1-p_i$

If one additional search is made, however Robbins (1968) noted that

$$V_1 = (n+1)^{-1} \sum_i I(X_i^{n+1} = 1). \qquad .....(4.3.3.)$$

is an unbiased estinator of θ_n. Robbins also argued that V_1 follows Un in the sense that the expected squared difference is strictly bounded from above by $(n+1)^{-1}$ Starr (1979) gave a more general version of the Robbins estimator. Starr supposed that the

initial search of size n was extended by m additional stage and defined

$$V_m = \sum_{k=1}^{m} \binom{m-1}{k-1}\binom{n+m}{k}^{-1} \sum_i I\,(X_i^{n+m} = k). \quad .....(4.3.4)$$

4.3.1. Properties of Starr's Estimator:

For convenience we begin by stating some results of Halmos (1946). A direct consequence of these results is the verification of Starr's conjecture. Another consequence is that V_m, defined in (4.3.4), is a U statistic. This property has further consequences, which we exploit.

To state Halmos's results, define II^* to be the class of all probability distribution on R, the real line. Let E be a Borel subset of R. Define II(E) to be the class of all $p \in II^*$ that have support in some finite subset of E, and let II be some subset of II^* that contains II (E). For each $P \in II$, Let $X_1, X_2, \ldots\ldots X_n$ be an i.i.d random sample. Let $\{i_1, \ldots i_k\}$ be a subset of size k of

$\{1,2,....N\}$ and let $\sum_c$ be the sum over all $\binom{N}{K}$ distinct combinations of $\{i_1,.....i_k\}$. A linear functional F(P) is said to be homogeneous of degree k if there exists a mapping h R^k to R such that

$$F(P) = E_p\ h\ (X_1\ ,\ .....X_k\ )$$

$$= \int\ .....\ \int h(x_1\ ,\ ..\ x_k\ )\ dP(x_1\ )....\ dP(x_k\ )\quad \text{for all}\quad p\in II$$

and if the integer k is minimal.

Lemma 4.3.1. (Halmos 1946, the theorems 3 and 5). Let F(P) be homogeneous of degree k over II with

$F(P) = E_p\ h(X_1, ...X_k)$.

1. If $f(X_1\ ,\ .....X_N)$ is a symmetric, unbiased estimate of F(P), then for every point $(x_1\ ,\ ...x_n)$ with $x_i\in E$, $f(x_1\ ,\ ....X_N)\ =$ $\binom{N}{k}^{-1}\ \sum_c\ h\ (X_{i1};\ ...X_{ik})$.

2. Among all unbiased estimator of F(P), $\binom{N}{k}^{-1} \sum_{c} h(X_{i_1}, \ldots . X_{i_k})$ has minimum variance.

To prove Starr's conjecture, define $E=\{1,2, \ldots\}$, $N=n+m$ and let II be the set of all probability distributions defined on E. We shall find the form of h(.) that is appropriate for this application. To motivate the discussion, we note that the indicator of the ith species having one representative can be expressed by $I(X_i^{n+1}=1)=\sum_{j=1}^{n+1} I(X_j=i) \prod_{\substack{k=1 \\ k\neq j}}^{n+1} I(X_k = i)$ (4.3.1.1).

We use the Kernel function of size n+1 defined by

$$h(X_1, \ldots . X_{n+1}) = (n+1)^{-1} \sum_{i} \sum_{j=1}^{n+1} I(X_j = i) x \prod_{\substack{k=1 \\ k\neq j}}^{n+1} I(X_k \neq i) \quad (4.3.1.2)$$

that is, the proportion of species with one representative. It is easy to see that h(.) is symmetric and unbiased for θ_n. The proof that θ_n is homogeneous of degree K=n+1 over I I standard and is given in Appendix A (Lemma A.1) Thus, by Lemma 4.3.1, we immediately have the following properties.

PROPERTY 4.3.1.1. The statistic V_m is a U statistic with Kernel h(.) and degree n+1; that is,

$$V_m \binom{n+m}{n+1}^{-1} \sum_c h\,(X_{i\,1}, \ldots\ldots X_{i\,n+1}) \qquad \ldots\ldots\ldots\ldots (4.3.1.3)$$

Property 4.3.1.2:- Based on a random sample of size n+m, V_m is the MVUE for θn over II.

A consequence of property 4.3.1.2 is that V_m has desirable property as an estimator of θn for any fixed number m additional searches. If the number of additional searches is large, from

Property 4.3.1.1 and the theory of U statistics it immediately follows that $V_m \longrightarrow \theta_n$ with probability 1 as $m \longrightarrow \infty$. Thus the estimator convergence to the parameter of interest. The rate of convergence can further be described by the following property.

Property 4.3.1.3:- Define

$$\sigma^2 = \sum_i p_i q_i^{2n-2} (np_i - q_i)^2 - \left(\sum_i p_i q_i^{n-1} (np_i - q_i)\right)^2$$

Then $V_m = \theta_n + (n+m)^{-1/2} \sigma Z + Op((n+m)^{-1/2})$ ..(4.3.1.4) as $m \longrightarrow \infty$

where Z is a standard normal random variable.

Remarks:- The proof of property 4.3.1.2 is standard in the theory of U statistics (see serfling 1980, p.192). One need only check the calculation of the asymptotic variance that is provided in (Lemma A.2). Perhaps the most interting aspect of property 4.3.1.3 is the fact that in the case of equal species probabilities, it can easily be shown that $\sigma = 0$. Indeed, by another application of U statistic theory in Results section 4.3.1. We have the following property.

Property 4.3.1.4: Suppose that $p_1 = p_2 \ldots . p_{1/\mu} = \mu$ for some $\mu > 0$. Then

$$V_m = \theta_n \; (n+m)^{-1} \binom{n+1}{2} \; \mu \; (1-\mu)^{n-2} (\mu - 2(n+1)^{-1})(\chi^2 - 1) + Op((n+m)^{-1}$$

as $m \longrightarrow \infty$, where χ^2 is a chi-squared random variable with 1df.

Thus the rate at which V_m approaches θ_n in the important special case of equal probabilities is of an order of magnitude different from that of the general case (with respect to weak convergence to a nondegenerate distribution). This characteristic is important, since a comparison of various alternative estimators in this special case can be misleading when drawing conclusions about their relative performance in the more general setup of unequal probabilities. In other situations, Starr (1979) Chao (1981), and Banerjee and Sinha (1985) used the equiprobable case as examples of their results. It should also be noted that the equiprobable cells model is unlikely to arise nature when sampling for species, although it arises naturally in the cataloging

problem of, for example, Harris (1959).

Proof of the Results in Section 4.3.1.

Lemma A.1. The parameter θ_n is hamogeneous over II and is of degree n+1.

Proof:- Using the kernel function h defined in (4.3.1.2), we have $E_p\ (h(x_1, \ldots . x_{n+1}) = \theta_n$, and thus θ_n is homogenous of degree $\leq$ n+1. We now suppose that θ_n is homogeneous of degree h and show that $h \geq n+1$. Thus assume that there exists $\phi\ (x_1 .. x_n)$ so that

$$\sum(1-q_i)q_i^n = E_p\ (\phi(X_1, \ldots . X_n)) \cdots (A.1)$$

for all $p \in II$ suppose that II_i is a β subset of II so that $p_q(1) = q_{/2}$, $p_q(2) = p_q(3) = (2-q)/4$ and $II_1 = \{Pq \in II,\ 0<q<1\}$. With the choice of p_q, the left side of (A.1) is a polynomial in q of degree n+1 and the right side is a polynomial in q of degree, say, $h_1 \leq h$. Since these polynomials must be of the same degree, we have $n+1 = h_1 \leq h$. Define $h_{1n}(X_1) = E(h(X_1, \ldots . X_{n+1})/X_1) - \theta_n$. The proof of property 4.3.1.3 is complete with $\sigma^2 = (n+1)^2$ Var $(h_{1n}(X_1))$ and the following Lemma.

Lemma A.2.

$$\text{Var}\,(h_{1n}(X_1)) = \sum_i (p_i-(n+1)^{-1})^2 p_i q_i^{2n-2} - (\theta_n-(1+n^{-1})^{-1}\theta_{n-1})^2$$

Proof:- Use (4.3.1.2) to get

$$E\,(h\,(X_1,\ldots,X_{n+1})/X_1) = (n+1)^{-1}\sum_i q_i^{n-1}\{np_i I\,(X_1\neq i)+q_i I(X_1=i)\}.$$

Thus, by rearranging terms,

$$h_{1n}(X_1) = \sum_i q_i^{n-1}\,(p_i-(n+1)^{-1})(p_i-I(X_1=i)).$$

Hence

$$E\,h_{1n}\,(X_1)^2 = \sum_i p_i q_i^{2n-1}\,(p_i-(n+1)^{-1})^2$$

$$-\sum_{i\neq j} p_i p_j q_i^{n-1}\,q_j^{n-1}(p_i-(p_i-(n+1)^{-1})(p_j-(n+1)^{-1}),$$

Which gives the result upon a rearrangement of terms.

To prove Property 4.3.1.4, we need to examine the properties of the following projection of h;

$$h_{2n}(X_1,X_2) = E(h(X_1,\ldots X_{n+1})/X_1,X_2)-h_{1n}(X_1)-h_{1n}(X_2)-\theta_n$$

$$=(1-\mu)^{n-2}(\mu-2(n+1)^{-1})\ (I(X_1=X_2)-\mu) \qquad (A.2)$$

To see (A.2), first note that it is easy to check that $\theta_n = (1-\mu)^n$ and $h_{1n}(x_1)=h_{1n}(x_2)=0$. Now, use (4.3.1.2) to get $E(h(X_1,\ldots X_{n+1})X_1,X_2)$

$$=(n+1)^{-1}(1-\mu)^{n-2}\sum_i \{(n-1)\mu I\ (X_1\neq i)I(X_2\neq i)$$

$$+(1-\mu)(I(X_1=i)I(X_2\neq i)+I(X_1\neq i)(X_2=i))\}$$

$$= (1-\mu)^{n-2}\{1-2\mu n/(n+1)+(\mu-2(n+1)^{-1})I(X_1=X_2)\}$$

after some algebra. Subtracting θ_n yields (A.2). The proof of property 4.3.1.4 is now an application of a result independently due to Gregory (1977) and Serfiling (1980,p192).

Proof of Property 4.3.1.4:- Let $K=(1-\mu)^{n-2}(\mu-(2(n+1)^{-1})$

so that $h_{2n}(x_1,x_2)=K(I(x_1=x_2)-\mu)$. It is immediate that

Var $(h_{2n}(X_1,X_2)=K^2\mu(1-\mu)>0$. Now, let g be an arbitrary, measurable function such that

$E(g(X))^2<\infty$ and let X, λ be real constants

The forms of g(.) and λ satisfying

$$\lambda_g(x)\ E=\{h_{2n}(x,X)_g(X)\}$$

$$=k\mu\{\sum_i I(x_i=i)_g(i)-E\ g(X)\}$$

$$=K\mu\{g(x)-E\ g\ (X)\}$$

are of two types. If Eg (X)$\neq$0, then g(X)=Eg(X)/(1-λ ku) is a constant ($\neq$0) and thus λ =0. If E g(X) =$\mu\sum_i g(i)=0$, then $\lambda=\pm\mu$. Thus, for example, by Serfling (1980, p.19). We have the result.

4.3.2. Alternative Nonparameteric Estimator:

Starr's estemator V_m is attractive computationally, since it is the linear combination of the "frequency of frequencies". and it has disirable theoretical properties, since it can be described

as a U statistic. Because it is derived from summary statistics, however, there may be some loss of information in a finite number of additional searches, in some sense. For example, if we set m=1, then from (4.3.3) we see that V_1 is the sample propertion of species, with one representative. Note that this estimator treats species with one representative. Note that this estimator treats species with 0,2,3...n+1 representatives equally. Motivated by these heuristic arguments, we introduce the following non-parametric estimator of Θn based on an initial sample size n and additional search m. Define $\hat{p}_i = (n+m)^{-1}\sum_{j=1}^{n+m} I(X_j=i)$ and $\hat{q}_i = 1-\hat{p}_i$;(i=1,2...). The NPMLE of θ_n is defined to be

$$\hat{\theta}_m = \sum_i \hat{p}_i \hat{q}_i^{\,n} \qquad \text{.....(4.3.2.1.)}$$

Unlike V_m, $\hat{\theta}_m$ is a biased estimator of θ_n. Since $(n+m)\hat{q}_i$ is a binomial random variable, it is straight forward to write out the bias explicitly as a linear combination of powers of q_i and Stirling numbers of the second kind. Finite sample properties

of $\hat{\theta}_m$ are further discussed in Section 4.3.3. Asymptotically (as $m \longrightarrow \infty$), $\hat{\theta}_m$ behaves similarly to V_m. By the strong law of large numbers, with probability 1, $\hat{q}_i \longrightarrow q_i$, and it is not hard to show that $\hat{\theta}_m \longrightarrow \theta_n$ with probability 1 as $m \longrightarrow \infty$. We also have the following two asymptotic properties.

Property 4.3.2.1.:- Let σ^2 be as defined in property 4.3.1.3 Then

$$\hat{\theta}_m = \theta_n + (n+m)^{-1/2}\sigma Z \; Op\left((n+m)^{-1/2}\right)$$

as $m \longrightarrow \infty$

Property 4.3.2.2:- Suppose that $p_1 = p_2 = \ldots = p_i/\mu = \mu$ for some $\mu > 0$. Then

$$\hat{\theta}_m = \theta_n + (n+m)^{-1}\binom{n+1}{2}(1-\mu)^{n-2}(\mu - 2(n+1)^{-1}) \times (\mu\,(\chi^2 - 1)) + (1-\mu) + Op\left((n+m)^{-1}\right)$$

as $m \longrightarrow \infty$

The proofs of properties 4.3.2.1 and 4.3.2.2. are in results section 4.3.2. comparing properties 4.3.1.3 and 4.3.2.1.,we see

that V_m and $\hat{\theta}_m$ are asymptotically equivalent to the first order [i.e., $(n+m)^{-1/2}$]. An advantage of the NPMLE $\hat{\theta}_m$ is that, since strong consistent estimators of q_i and hence σ can be constructed, we have as an immediate corollary of property 4.3.2.1 large sample interval estimates of θ_n.Comparing properties 4.3.1.4 and 4.3.2.2, we see that V_m and $\hat{\theta}_m$ are of the same order of magnitude and have the same variance in their respective asymptotic distributions. The estimator V_m is slightly superior to $\hat{\theta}_m$ in the sense that the asymptotic distribution of $V_m-\theta_n$ has mean 0, unlike that of $\hat{\theta}_m-\theta_n$. We remark that in this special case of equiprobable cells, Chao's (1981) extension of Harris's (1968) estimator is MVUE for fixed m and hence is a strong competitor of V_m and $\hat{\theta}_m$.

As noted, the rate of convergence of V_m and $\hat{\theta}_m$ is markedly different in the equiprobable case in comparison with the general case. Moreover in some sense the equiprobable case is the only one in which this can happen. Specially we have the following result.

Property 4.3.2.3. Consider σ^2 defined in property 4.3.1.3 and

suppose that the number of species exceeds n. Then $\sigma^2 = 0$ iff $p_i = p_2 \cdots p_{1/\mu}$ for some $\mu > 0$.

Proof of the results in section 4.3.2.

Proof of property 4.3.2.1. Define $G(x)=X(1-X)^n$ and note that $\theta_n = \sum_i G(p_i)$ and that $\hat{\theta}_m = \sum_i G(p_i)$. By a Taylor - series expansion,

$$\hat{\theta}_m = \theta_n \sum_i (\hat{p}_i - p_i) G'(p_i) + O(\sum_i (\hat{p}_i - p_i)^2),$$

since $G''(x)$ is bounded for $0<x<1$. Now, since

$$(n+m)^{1/2} E\sum_i (\hat{p}_i - p_i)^2 = (n+m)^{-1/2} \sum_i p_i q_i \leq (n+m)^{-1/2} \longrightarrow 0 \qquad (B.1)$$

We have that $(n+m)^{1/2} \sum_i (\hat{p}_i - p_i)^2 \longrightarrow 0$ in probability.

By Fubini's theorem, we have that

$$\sum_i (\hat{p}_i - p_i) G'(p_i) = (n+m)^{-1} \sum_{j=1}^{n+m} \left\{ \sum_i G'(p_i)(I(X_j = i) - p_i) \right\}.$$

This, the central limit theorem, and Slustsky's theorem give the result.

Proof of Property 4.3.2.2. By a Taylor -series expansions

$$\hat{\theta}_m = \theta_n + G''(\mu)/2 \sum_i (\hat{p}_i - \mu)^2 + 0(\sum_i (\hat{p}_i - \mu)^3),$$

since $G'''(x)$ is bounded for $0 < x < 1$ and $\sum_i (\hat{p}_i - \mu) = 0$.

Similarly to (B.1) we have that $(n+m) \sum_i (\hat{p}_i - \mu)^2 \longrightarrow 0$.

in probability. Thus

$$(n+m)(\hat{\theta}_m - \theta_n) = (n+m)G''(\mu)/2(\sum_i \hat{p}^2 - \mu^2) + O_p(1) \qquad \ldots\ (B.2)$$

Now

$$\sum_i (\hat{p}_i - \mu^2) = \sum_i \left((n+m)^{-1} \sum_{j=1}^{n+m} I(X_j = i) \right)^2 - \mu$$

$$= (n+m)^{-1} + 2(n+m)^{-2} \sum_{j<k} I(X_j = X_k) - \mu$$

$$= (n+m)^{-1}(1-\mu) + (1-(n+m)^{-1})U, \qquad \ldots\ldots\ (B.3)$$

Where $U = \binom{n+m}{2}^{-1} \sum_{j<k} I(X_j = X_k) - \mu$ is a U statistic.

As in the proof of poropety 4.3.1.4, $E(U/X_1)=0$ and

$$\binom{n+m}{2} E(U/X_1, X_2) = (I(X_1 = X_2) - \mu).$$

Thus, by the same argument as in the proof of property 4.3.1.4 (with k=1), we have

$$(n+m)\ U \longrightarrow \mu(\chi^2 - 1)$$

This, (B.2), B.3), and Slutsky's theorem yield the result.

Proof of property 4.3.2.3.:- We need only show that $\sigma^2=0$ implies that $p_i = p_j$ for each i,j. To do this we construct the random variable $X=(np_i - q_i)q_i^{n-1}$ with probability p_i (i=1,2,....). Now it is easy to see that $Var(X)=\sigma^2$ and thus $\sigma^2=0$ means that $(np_i - q_i)q_i^{n-1} = (n+1)(p_i - (n+1)^{-1})q_i^{n-1}$ must be some constant C for i=1,2, ... Since the number of species exceeds n, we have $p_i \leq (n+1)^{-1}$ for some i and C must be nonpositive. The question of whether

different pi may satisfy $(n+1)(p_i-(n+1)^{-1})q_i^{n-1} = C$ is equivalent to finding the number of roots of $h(x) = (n/(n+1)-x)x^{n-1}-C$, $0 < x < 1$.

Now, $h'(x)=x^{n-2}((n-1)/(n+1)-x)$ is positive for $0<x<(n-1)/(n+1)$ and is negative for $(n-1)/(n+1)<x\leq 1$. Further, $h(0)=-C$ and $h(1)=-(n+1)^{-1}-C$. Thus for $-(n+1)^{-1}<C<0$ there is exactly one root and for $C<-(n+1)^{-1}$ there are no roots

4.3.3. Small Sample Properties:

In this section we investigate, via a Monte Carlo simulation, the behavior of Starr's estimator, V_m, and the NPMLE, $\hat{\theta}_m$, when m is small. We look at their bias and mean squared error as estimates of θ_n and make some comments regarding modifications of $\hat{\theta}_m$ that have desirable properties. Finally, we investigate modifications of V_m and $\hat{\theta}_m$ suitable for use when m=0. All computations were done on a VAX11/750 owned and operated by the Department of Statistics at the University of Wisconsin - Madison. The simulations were performed using the National Bureau of Standards's Core Math Library (CMLIB) pseudouniform random number generator UNI.

Two classes of distributions were used to construct the probability distribution $\{p_i; i \geq 1\}$. These were (a) equiprobable, with $p_i = \mu$ $(1 \leq i \leq 1/\mu)$, and (b) truncated geometric with $p_i = qp^{i-1}/(1-p^c)(1 \leq i \leq c; 0<p<1;\ q=1-p)$. For the equiprobable cells model, values of μ =1,.02,.01 were used; for the truncated geometric model, values of p=.1, .5,.9 and C=10,100 were used. For each assignment of $\{p_i\}$, θ_n was determined and 1,000 simulations were performed. For each simulation, this involved drawing a sample of size n and a subsequent sample of size m. The pairs (n,m) = (10,1), (10,10),(50,1), (50,10), (50,50) were included. For each sample, $\hat{\theta}_m$ and V_m were computed. Tables 1 and 2 show the mean values of $\hat{\theta}_m$ and V_m over the 1000 sample denoted in the tables by $E\hat{\theta}_m$ and E V_m respectively. (The rows corresponding to m=0 will be discussed later) In addition, the estimated root mean squared error of the estimates, denoted by RMSE $(\hat{\theta}_m)$ and RMSE (V_m), respectively, are given in Tables 1 and 2. Of course, since V_m is unbiased, RMSE (V_m) is also an estimate of the standard error of V_m.

Generally, in the equiprobable case, Vm has lower RMSE than $\hat{\theta}_m$. Comparing propeties 4.3.1.4 and 4.3.2.2 we have up to order $(n+m)^{-1}$, $E\,(V_m-\theta_n)^2 = \dfrac{E\,(\hat{\theta}_m-\theta_n)\;2\mu}{(2\mu^2-2\mu+1)}$ Thus, for μ small, RMSE (V_m) will be approximately $2\mu^2$ times RMSE $(\hat{\theta}_m)$. Although the difference in RMSE for V_m andf $\hat{\theta}_m$ in Table 1 are not all of this magnitude, we do see that V_m is a better estimator of $\hat{\theta}_m$ in terms of RMSE.

The situation is reversed to a large extent when the truncated geometric is used for $\{p_i\}$. These results appear in Table 2. It is evident in this case, as in Table 1, that $\hat{\theta}_m$ tends to underestimate θ_n and that the bias can be considerable.From the results of Section 4.3.1. & 4.3.2, we expect $\hat{\theta}_m$ and V_m to have the same asymptotic mean squared error. From Table 2, it appears that when p is note too large, the mean squard error of $\hat{\theta}_m$ is less than V_m, some times considerably so. That this can fail when p is large is not suprising, since the truncated geometric distribution tends

to the terms of its mean squared error, it has already been noted that its bias can be considerable. In fact,

$$E(\hat{\theta}_m) = \theta_n + (n+m)^{-1}\left\{\binom{n}{2}\theta_{n-1} - \binom{n+1}{2}\theta_n\right\} + o((n+m)^{-1})$$

This suggests that the quantity

$$\hat{\theta}_m + (n+m)^{-1}\left(\binom{n+1}{2}\hat{\theta}_m - \binom{n}{2}\sum_i \hat{p}_i \hat{q}_i^{n-1}\right)$$

would be a better estimator of θ than $\hat{\theta}$ alone. For the size of the samples discussed here $\sum_i \hat{p}_i \hat{q}_i^{n-1}$ tends to underestimate θ_{n-1} too severely and a better estimator can be obtained by replacing $\sum \hat{p}_i \hat{q}_i^{n-1}$ by $\hat{\theta}_m$, leading to the estimator.

$$\theta_m^* = \hat{\theta}_m(1 + n/(n+m)). \qquad (4.3.3.1)$$

Values of $E(\theta_m^*)$ and RMSE (θ_m^*) are given in Tables 1 and 2. Generally, θ_m^* has good bias properties and compares favorably with V_m interms of RMSE, even for the equiprobable case.

It should be noted that $\hat{\theta}_m$ and V_m are, in some sense, "retrodictors". That is they predict, on the basis of n+m observations, what would be observed for the last m observations. In Starr (1979), an argument is given that this is not a vacuous

exercise; V_m can be used effectively to predict, on the basis of an initial sample size n and a subsequent sample of size m, what will occur in a large future sample of size M. This argument applies equally well to the NPMLE $\hat{\theta}_m$. It can be argued, however, that the to the equiprobable case when p tends to 1.

Specically,

$$qp_i/(1-p^c) \longrightarrow 1/c \text{ for each i as } p \longrightarrow 1.$$

That $\hat{\theta}_m$ dominates V_m in terms of the truncated distribution when p is small can be seen in an example as follows. Let C=2, m=1 and n=2, so $p_1 = q(1-p^2)$ and $p_2 = qp/(1-p^2)$. Then $\theta_2 = p_1 p_2$ and it is easy to show that $E\ \hat{\theta}_1 = \frac{2}{3}\ p_1 p_2$, which represents a considerable bias. For this example it can be shown that $E(\hat{\theta}_1 - \theta_2)^2 = (4p_1p_2 - 9p_1^2p_2^2)/27$ and $E(V_1 - \theta_1)^2 = p_1p_2\ (p_1 - p_2)^2$. It follows that

$E(V_1 - \theta_2)^2 \geq E\ (\hat{\theta}_1 - \theta_2)^2$ if $p_2 \leq \frac{1}{2} - \sqrt{77/66}$, or equivalently, if $p \leq .5799$.

Although $\hat{\theta}_m$ may be an attractive estimator in the truncated geometric case in principal interest of estimators such as $\hat{\theta}_m$ and V_m is in their properties as true predictors. For example, Rasmussen and Starr (1979) used the estimator $V_o = n^{-1}\sum_i^n I\ (X_i = 1)$ to consider a rule for subquentially. Sampling a population. Similarly, the estimators $\hat{\theta}$ and θ^* could also be used in such a capacity. We leave the examination of such sequential rules to a

future paper and consider here only the properties of V_o; $\hat{\theta}_o$ and θ_o^* estimates of θ_n. Simulation results appears in Tables 1 and 2. In terms of mean squared error, again we see that, in the equiprobable case, Vodominates $\hat{\theta}_o$ and that θ_o^* compares favorably with V_o. In the truncated geometric case, both $\hat{\theta}_o$ and θ_o^* dominate V_o, except when p is near 1, in which case V_o tends to be better estimator than $\hat{\theta}_o$.

4.3.4. Summary and Discussion:

This article has focused on nonparameteric estimators of θ_n the probability of discovering a new species. We have shown V_m to be an MVUE with a high rate of convergence in the equiprobable case. The non-parametric maximum liklihood estimator, $\hat{\theta}_m$ has similar asymptotic properties. In small samples, V_m is a better estimator than $\hat{\theta}_m$ in the equiprobable cell case with respect to mean squared error; this is reversed for truncated geometric distributions when p is not large. An estimator with some what less bias than $\hat{\theta}_m$ is θ_m^*, defined in (4.3.3.1); it compares favorably with V_m in terms of mean squared error.

Besides the theoretical interest in $\hat{\theta}_m$ as an estimator that competes well with V_m in the truncated geometric case, we argue that this has practical implications. For example, data collected by J. Andrews (personal communication 1985) of the species

abundance of epiphytic fungi on apple leaves fit a truncated geometric distribution quite well with p=.77. Further arguments were given by Pielou (1975), stating that a geometric distribution is appropriate in some situations for modeling species distributions. In a future study the comparison of V_m and $\hat{\theta}_m$ over a wider class of distributions will be addressed. We conjecture that $\hat{\theta}_m$ will dominate V_m whenever the underlying distribution $\langle p_i ; i \geq 1 \rangle$ is sufficiently nonuniform.

Table 1 Equiprobable Case

μ	n	m	θ_n	EV_m	$RMSE(V_m)$	$E(\hat{\theta}_m)$	$RMSE(\hat{\theta}_n)$	$E(\hat{\theta}_m)$	$RMSE(\theta^*_m)$
.1	10	0	.3487	.3923	.1660	.1836	.1717	.3671	.0961
	10	1	.3487	.3488	.1364	.1951	.1603	.3724	.0904
	10	10	.3487	.3471	.0555	.2598	.0555	.3897	.0658
	50	0	.0052	.0057	.0099	.0136	.0953	.0271	.0231
	50	1	.0052	.0058	.0101	.0136	.0092	.0269	.0229
	50	10	.0052	.0053	.0060	.0125	.0080	.0230	.0187
	50	50	.0052	.0050	.0025	.0097	.0050	.0146	.0099
.02	10	0	.8171	.8376	.1577	.3085	.5102	.6169	.2153
	10	1	.8171	.8242	.1469	.3403	.4783	.6496	.1828
	10	10	.8171	.8183	.0782	.5108	.3088	.7661	.0775
	50	0	.3642	.3716	.0695	.1932	.1723	.3863	.0470
	50	1	.3642	.3656	.0670	.1959	.1695	.3880	.0479
	50	10	.3642	.3618	.0530	.2168	.1488	.3975	.0509
	50	50	.3642	.3639	.0260	.2731	.0925	.4097	.0515
.01	10	0	.9044	.9146	.1240	.3280	.5772	.6560	.2556
	10	1	.9044	.9051	.1206	.3614	.5439	.6900	.2224
	10	10	.9044	.9054	.1550	.5534	.3522	.8301	.0869
	50	0	.6050	.6101	.0818	.2650	.3407	.5300	.0870
	50	1	.6050	.6047	.0799	.2699	.3359	.5344	.0829
	50	10	.6050	.6033	.0690	.3078	.2981	.5643	.0591
	50	50	.6050	.6059	.0370	.4277	.1983	.6116	.0308

Table 2 Truncated Geometric Distribution

p	c	n	m	θ_n	EV_m	$RMSE(V_m)$	$E\hat{\theta}_m$	$RMSE(\hat{\theta}_m)$	$E\theta^*_m$	$RMSE(\theta^*_m)$
.1	10	10	0	.0443	.0494	.0566	.0241	.0293	.0249	.0367
		10	1	.0443	.0457	.0530	.0230	.0383	.0439	.0357
		10	10	.0443	.0460	.0274	.0324	.0204	.0486	.0252
		50	0	.0075	.0078	.0108	.0047	.0051	.0093	.0087
		50	1	.0075	.0079	.0108	.0047	.0050	.0094	.0085
		50	10	.0075	.0071	.0087	.0047	.0051	.0086	.0078
		50	50	.0075	.0072	.0058	.0055	.0044	.0083	.0060
.5	10	10	0	.1302	.1454	.1036	.0708	.0691	.1416	.7117
		10	1	.1302	.1308	.0950	.0745	.0659	.1423	.0686
		10	10	.1302	.1292	.0515	.0968	.0467	.1452	.0512
		50	0	.0273	.0274	.0200	.0158	.0136	.0315	.0148
		50	1	.0273	.0286	.0189	.0160	.0132	.0317	.0141
		50	10	.0273	.0265	.0162	.0173	.0123	.0317	.0136
		50	50	.0273	.0274	.0103	.0211	.0091	.0317	.0109
.9	10	10	0	.3319	.3648	.1567	.1731	.1655	.3463	.0950
		10	1	.3319	.3145	.1386	.1876	.1515	.3582	.0920
		10	10	.3319	.3314	.0604	.2487	.0914	.3730	.0702
		50	0	.0096	.0107	.0131	.0161	.0078	.0322	.0242
		50	1	.0096	.0098	.0130	.0160	.0077	.0317	.0237
		50	10	.0096	.0097	.0083	.0156	.0073	.0286	.0205
		50	50	.0096	.0098	.0042	.0139	.0053	.0208	.0122

.1	100	10	0	.0443	.0494	.0566	.0214	.0293	.0428	.0367
		10	1	.0443	.0388	.0498	.0208	.0297	.0396	.0348
		10	10	.0443	.0438	.0270	.0310	.0213	.0465	.0251
		50	0	.0075	.0078	.0108	.0047	.0051	.0093	.0087
		50	1	.0075	.0075	.0106	.0045	.0051	.0089	.0083
		50	10	.0075	.0077	.0093	.0049	.0051	.0090	.0081
		50	50	.0075	.0074	.0057	.0057	.0042	.0086	.0058
.5	100	10	0	.1312	.1471	.1057	.0719	.0691	.1438	.0720
		10	1	.1312	.1305	.0951	.0759	.0656	.1448	.0686
		10	10	.1312	.1318	.0543	.0982	.0476	.1473	.0539
		50	0	.0283	.0290	.0212	.0163	.0141	.0327	.0157
		50	1	.0283	.0277	.0205	.0163	.0142	.0322	.0152
		50	10	.0283	.0284	.0163	.0180	.0126	.0329	.0140
		50	50	.0283	.0291	.0108	.0222	.0093	.0333	.0116
.9	100	10	0	.6095	.0269	.1890	.2505	.3628	.5011	.1511
		10	1	.6095	.6150	.1792	.2771	.3368	.5220	.1319
		10	10	.6095	.6084	.1051	.4011	.2161	.6017	.0861
		50	0	.1855	.1856	.0533	.1030	.0846	.2060	.0428
		50	1	.1855	.1856	.0509	.1059	.0816	.2098	.0435
		50	10	.1855	.1857	.0447	.1155	.0725	.2117	.0437
		50	50	.1855	.1857	.0269	.1410	.0477	.2115	.0367

REFERENCES

REFERENCES

1. BANERJEE, P.K. and SINHA, B.K. (1985),"Optimal and Adaptive strategies in Discovering New Species," Sequential Analysis, 4, 111-122.

2. CHAO, A. (1982) Correction to "On Estimating the probability of Discovering a New species, "The annals of Statistics, 10,1311.

3. CHAO, A. (1984), "Nonparametric Estimation of the Number of classes in a population "Scandinavian Journal of Statistics" 11, 265-270.

4. EFRON, B.and Thisted R, (1976),"Estimating the Number of Unseen species, Biometrika, 63, 435-447.

5. GOOD, I. (1953), "On the population frequencies of species and the estimation of population parameters, "Biometrika, 40, 237-264.

6. HILL, BRUCE M. (1968), "Posterior Distribution of Percentiles Baye's Theorem for Sampling from a population", Journal of the American Statistical Association, 63, 677-691.

7. I.P. GERASIMOV (1981) Geography and Ecology 'Progress Publishers Moscow.

8. LUDWIG, JOHN A. REYNOLDS, JAMES F. (1988) "STATISTICAL ECOLOGY" (A primer on methods and computing) JohnWiley & Sons New York.

9. MANLY,BRYAN F.J. (1985) The statistics of Natural selection on Animal populations CHAPMAN AND HALL Ltd. LONDON.

10. MURRAY K. CLAYTON and EDWARD W. FREES (1987) "Nonparametric Estimation of the Probability of Discovering a New species" Journal of the American Statistical Association 82, 305-311.

11. NELSON, G. HAIRSTON, SR. (1989) "Ecological experiments (Purpose, design and execution) Cambridge University Press Cambridge.

12. PIELOU, E.C. (1975), Ecological Diversity, New York: John Wiley & Sons, New York.

13. PIELOU, E.C. (1977) Mathematical Ecology JOHN WILEY & SONS, New York.

14. PIELOU, E.C.(1984) The interpretation of Ecological data. John Wiley & Sons New York.

15. Yu M. SVIREZHEV, D.O. LOGOFET (1983) Stability of Biological Communities, Mir Publishers Moscow.